

To the Graduate Council:

I am submitting herewith a dissertation written by Jalel Zrida entitled "State Estimation in Stochastic Decision Theory." I have examined the final copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Electrical Engineering.

J. Douglas Birdwell
J. Douglas Birdwell, Major Professor

We have read this dissertation
and recommend its acceptance:

J. R. B. Cockett

J. M. George

J. A. Lawrence

S. Samuel Jordan

James H. K.

Accepted for the Council:

Lawrence Minkel
Vice Provost
and Dean of The Graduate School

STATE ESTIMATION IN STOCHASTIC DECISION
THEORY

A Dissertation

Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Jalel Zrida

December 1988

Copyright © Jalel Zrida, 1988
All rights reserved

DEDICATION

To my father Mansour Zrida, my mother Halima Hafsi,
my sisters Radhia and Mona, and
to my brothers Hichem, Kamel, and Mehdi.

ACKNOWLEDGEMENTS

In achieving the results in my research, two professors were essential to my progress: to Dr. J. Douglas Birdwell, my thesis advisor, and Dr. J. Robin B. Cockett, I am particularly indebted for their consultation and criticism. The remaining members of my committee, Dr. Jack Lawler, Dr. Joseph Googe, Dr. Sam Jordan, and Dr. James K. Ho greatly contributed to my efforts by their comments and encouragement. I gratefully acknowledge the beneficial effects of their participation in my endeavors. I also would like to salute the "215 gang" at Ferris Hall, especially Dimitrios Bargiotas, Gregory Murphy, Guiran Chang, and Jim Patton and thank them for their friendship and help. I also would like to thank the University Mission of Tunisia for the support and assistance they provided me during the course of my studies in the United States.

This research was partially supported by the Oak Ridge National Laboratory, Power Systems Technology Program, under Subcontract Nos. 41B-07685-C/S89 and 41B-07685-C/T85, and by the National Science Foundation under Grant No. ECS-8715092.

ABSTRACT

This dissertation addresses the problem of state estimation by viewing it as a decision-making process under uncertainty. An efficient structure for knowledge representation, in an uncertain environment, is introduced in the framework of *stochastic decision theory*. Compared to some alternate representations, this structure unifies the uncertain character of the knowledge with the representation. Furthermore, this structure is exclusively armed with manipulation capabilities which regulate information storage and retrieval.

The problem of incomplete information is also addressed in the context of stochastic decision theory using the maximum entropy principle. A second important problem which arises in any uncertain knowledge representation structure is the problem of statistical independence/dependence between the propositions of the knowledge. The dissertation provides an algorithm which simultaneously uncovers, for a given stochastic term, all the independence and dependence classes.

The dissertation explores how this structure can be used to approach some classical problems in *state estimation*. This provides a new visualization of the state estimation procedure in terms of syntactic manipulation. The results of the entailment and the statistical independence problem are used to devise a cluster decomposition scheme for stochastic terms, similar to machine decomposition in automata theory, which allows the decomposition of the state estimation problem into smaller sub-problems. This finding helps avoid the crucial storage and computational explosion encountered in large scale systems.

TABLE OF CONTENTS

CHAPTER	PAGE
1. INTRODUCTION	1
2. RELEVANT LITERATURE	12
2.1 STATE ESTIMATION THEORY	13
2.2 NILSSON'S SEMANTIC TREES	20
2.3 PEARL'S BELIEF NETWORKS	22
2.4 THE STOCHASTIC DECISION THEORY REPRESENTATION	25
3. BACKGROUND	32
3.1 CATEGORY THEORY	32
3.2 DISCRETE DECISION THEORY	42
3.3 MEASURE AND PROBABILITY THEORY	55
3.3.1 Definitions in Measure Theory	55
3.3.2 Probability Theory	56
3.4 SUMMARY	62
4. STOCHASTIC DECISION THEORY	63
4.1 INTRODUCTION	63
4.2 THE SETTING	64
4.3 STOCHASTIC DECISION THEORIES	71
4.4 MODELS OF STOCHASTIC DECISION THEORIES	81
4.5 STRUCTURAL RETRIEVAL AND EQUIVALENCE	84

4.6	GENERAL RETRIEVAL AND COMPLETENESS	93
4.7	SUMMARY AND CONCLUSION	97
5.	PROBABILISTIC ENTAILMENT IN STOCHASTIC DECISION THE- ORY	98
5.1	ENTROPY IN STOCHASTIC DECISION THEORY	99
5.2	THE MAXIMUM ENTROPY PRINCIPLE	100
5.2.1	The Unconstrained Case	102
5.2.2	The Constrained Case	102
5.3	AN APPLICATION EXAMPLE	103
5.4	CONCLUDING REMARKS	106
6.	STATISTICAL INDEPENDENCE IN STOCHASTIC DECISION THE- ORY	107
6.1	STATISTICAL INDEPENDENCE	108
6.2	DEPENDENCE/INDEPENDENCE GENERATING ALGORITHM	110
6.2.1	Preliminaries	111
6.2.2	Statistical Independence	112
6.2.3	Generating Dependence/Independence Classes	113
6.3	EXAMPLE	116
6.4	CONCLUSION	119
7.	A STATE ESTIMATION PROBLEM	120
7.1	A CLASSICAL STATE ESTIMATION PROBLEM	121
7.2	STATE ESTIMATION IN LARGE SCALE SYSTEMS	123
7.3	THE STOCHASTIC DECISION THEORY APPROACH	124

7.3.1	The Insight of Stochastic Decision Theory	124
7.3.2	An Example	125
7.4	A CLUSTER DECOMPOSITION SCHEME	130
7.4.1	Objective	131
7.4.2	Probabilistic Entailment	132
7.4.3	Statistical Dependence	133
7.4.4	The Approximation Method	135
7.5	THE CLUSTER DECOMPOSITION AND THE STATE ESTIMATION PROBLEM	140
7.5.1	Decomposition of the State estimation Problem	140
7.5.2	An Example	141
7.6	CONCLUDING REMARKS	143
8.	CONCLUSIONS AND FUTURE DIRECTIONS	145
	BIBLIOGRAPHY	148
	APPENDICES	154
A.	APPENDIX TO CHAPTER 3	155
A.1	APPENDIX TO SECTION 3.3	155
A.1.1	Proof of Theorem 4 on Page 59	155
B.	APPENDIX TO CHAPTER 4	157
B.1	APPENDIX TO SECTION 4.2	157
B.1.1	Proof of Lemma 6 on Page 70	157
B.1.2	Proof of Lemma 7 on Page 71	157

B.2	APPENDIX TO SECTION 4.3	158
B.2.1	Proof of Lemma 8 on Page 72	158
B.2.2	Proof of Theorem 6 on Page 77	159
B.2.3	Proof of Theorem 7 on Page 78	160
B.2.4	Proof of Corollary 1 on Page 79	165
B.3	APPENDIX TO SECTION 4.4	167
B.3.1	Proof of Theorem 8 on Page 82	167
B.3.2	Proof of Theorem 9 on Page 83	170
B.4	APPENDIX TO SECTION 4.5	172
B.4.1	Proof of Lemma 9 on Page 85	172
B.4.2	Proof of Theorem 10 on Page 87	172
B.4.3	Proof of Lemma 10 on Page 91	173
B.4.4	Proof of Lemma 11 on Page 92	174
B.4.5	Proof of Theorem 11 on Page 92	175
B.5	APPENDIX TO SECTION 4.6	176
B.5.1	Proof of Theorem 12 on Page 95	176
B.5.2	Proof of Lemma 12 on Page 96	179
B.5.3	Proof of Theorem 13 on Page 97	180
C.	APPENDIX TO CHAPTER 5	181
C.1	APPENDIX TO SECTION 5.1	181
C.1.1	Proof of Lemma 13 on Page 99	181
C.1.2	Proof of Lemma 14 on Page 100	181
C.1.3	Proof of Corollary 2 on Page 100	182
C.2	APPENDIX TO SECTION 5.2	183

C.2.1	Proof of Theorem 14 on Page 102	183
C.2.2	Proof of Theorem 15 on Page 103	184
D.	APPENDIX TO CHAPTER 6	186
D.1	APPENDIX TO SECTION 6.1	186
D.1.1	Proof of Theorem 16 on Page 109	186
D.1.2	Proof of Corollary 3 on Page 110	188
D.2	APPENDIX TO SECTION 6.2	188
D.2.1	Proof of Lemma 15 on Page 111	188
E.	APPENDIX TO CHAPTER 7	190
E.1	APPENDIX TO SECTION 7.3	190
E.1.1	Proof of Theorem 17 on Page 125	190
E.2	APPENDIX TO SECTION 7.4	191
E.2.1	Proof of Lemma 18 on Page 132	191
E.2.2	Proof of Lemma 19 on Page 133	192
E.2.3	Proof of Theorem 18 on Page 138	193
VITA		194

LIST OF FIGURES

FIGURE	PAGE
2.1 Nilsson's Representation.	21
2.2 Stochastic Decision Theory Construction.	27
2.3 Stochastic Decision Theory Representation.	27
2.4 Stochastic Decision Tree Representation.	28
2.5 Transposed Stochastic Decision Tree.	30
2.6 The Augmented Decision Tree.	30
2.7 The Updated Decision Tree.	31
4.1 The Stochastic Decision Tree t of the Retrieval Example.	90
4.2 The Stochastic Decision Tree t' Under the Structural Retrieval Procedure.	91
5.1 The 4-Quadrant Non-Moving Target Problem.	104
5.2 The Stochastic Term t_1	104
5.3 The Stochastic Term t_2	105
5.4 The Stochastic Term t_3	106
6.1 The Stochastic Term t_1	116
6.2 The Stochastic Term t_2	118
6.3 The Stochastic Term t_3	118
7.1 The Example's Stochastic Decision Tree.	126
7.2 The Augmented Stochastic Decision Tree.	128
7.3 The Augmented Decision Tree after the Estimation Procedure.	129
7.4 The Updated Stochastic Decision Tree of the Example.	129
7.5 The Already Augmented Stochastic Term.	142

7.6	The Terms of the Cluster Decomposition.	143
7.7	The Cluster Decomposition Terms After the Estimation Procedure.	144

CHAPTER 1

INTRODUCTION

Physical systems are designed to perform in a usually well-defined way. However, without further precautions, this is rarely the case. As a result, one should be, at any time, monitoring the system in operation to make sure that it is performing correctly according to specifications. In other words, one must know the “state” of the system at any time. The concept of *state* of a system, at a given time t_0 , is defined, in this context, as the amount of information about the system at t_0 which, together with all the inputs, if any, applied to the system since time t_0 , determine uniquely the behavior of the system [11].

Because physical systems are always inaccurately modeled and always subject to random disturbances, the state is always random. One must therefore take *measurements* or *observations* on the system in order to determine its state. These observations are in turn contaminated by noise caused by the sensors or measuring devices. *State estimation* is the process of extracting the probable state of a system from corrupted measurements taken on the system. This estimation process uses known relationships to infer the desired information from the measurements, taking into account measurement errors, the dynamics of the system, and prior knowledge about the system.

The problem of estimation dates back almost two centuries to the work of Gauss and still is of central importance in control systems theory. In order to observe a system, control it, or optimize its performance, estimates of the state are always necessary. The importance of the problem has led to a wide variety of solutions. The least squares tech-

nique approaches the problem from a statistical point of view. Difficulties in interpreting the least squares solution suggested a probabilistic approach which not only rederived the least squares solution, but also led to interesting results in both the linear and nonlinear case. The minimum variance filter, the maximum likelihood filter, and the Kalman filter are some of these celebrated results that broaden considerably the scope of state estimation theory. The implementation of these filters, however, faces a serious problem of computational and storage burden especially with the ever-increasing size of contemporary applications. This has led many authors to settle for suboptimal filters; this dissertation is another endeavor in this direction. This endeavor is, however, unique in the sense that the estimation problem is looked upon from a different perspective: The stochastic decision theory perspective views the state estimation problem as a decision-making process under uncertainty. This theoretic decision setting requires a representation tool for uncertain knowledge similar to the one used in expert system technology. In fact, many engineering problems can be viewed, in this context, as an attempt to correct, update, or upgrade a current state of *knowledge* specific to a given domain. We assume that this knowledge not only reflects the current information about the solution of the problem, but also how to extract additional information from the system environment such as information derived from analysis or design tools, measurement devices, or humans. We also assume that this knowledge is incomplete, corrupted, and not known with certainty. The updated knowledge will be in turn scrutinized for further updating, and the process goes on.

The problem which naturally arises is how to make knowledge available in an efficient manner which not only can cope with the large size of contemporary problems, but also remains easy to manipulate. The size problem has been reduced by the continuing increase in computational power per unit price. How well knowledge is represented with respect to handling and manipulation is another matter. The nature of the problem and the

resources that are available restrict the capabilities of the representation structure.

This constitutes the starting point of this dissertation in which structures similar to the ones used in expert systems are considered and a suitable representation tool is adopted which not only deals with uncertainty but also with manipulating operations such as the one involved in expert systems (*e.g.*, information retrieval and storage).

The development of intelligent software or expert systems, which contain human-like reasoning abilities, has emerged as one of today's most exciting research areas. These expert systems are an attempt to integrate human knowledge into a computerized structure, called a knowledge base. This structure constitutes the production system, which is a collection of facts and rules describing the knowledge. Each piece of this production system is referred to as a production rule. The most common pattern of production rule is the following:

condition IMPLIES action.

The condition is usually one or more tests on the current state of facts; the result triggers the action, if the tests are satisfied. The action may produce conclusions, or alter the current state of facts, or both.

A production system consists of a set of condition/action production rules operating in cycles. During each cycle, the conditions of each production rule are matched against the current state of facts. When the current fact matches the condition in any production rule, the action in that production rule is taken. This action usually changes the current state of facts, producing new facts that will be subsequently tested. After each cycle, a stopping condition is checked which, when it is true, terminates the process. In this way, expert systems have been able to provide a reasonable simulation of human decision making with efficiency and repeatability.

Decision processes can be well represented by means of decision trees. Decision trees are a sequential evaluation procedure of concurrent processes where, at each step, a given process is evaluated and the next action is chosen accordingly. Because decision trees are a primitive and intuitive representation tool, they have been considered in a wide variety of disciplines including switching theory, pattern recognition [22], machine diagnosis, and expert systems [59,60]. They have also been criticized for becoming quickly inefficient and hard to read and manage with increasing problem size [21]. As a result, rule-based methods ("if-then-else rules") for knowledge representation gained more popularity. The reasons for this popularity stems perhaps from the fact that this representation is easy to establish and to maintain and is compatible with logic programming languages such as Prolog and Lisp. This fact and the availability of "inference engine" capabilities have shifted attention from the important issue of complexity of rule manipulation (almost all significant manipulations of rules are NP-complete) [19]. In [19], the authors addressed the so-called control problem of rule-based representations. A pure backward (or forward) chaining system cannot be used, always, to control the order in which rules are fired. This problem has led to the introduction of meta-rules, a higher-level set of rules which control the order of rule application. These meta-rules, in turn, need meta-meta-rules themselves, and the problem remains unsolved. Decision trees do not have this problem, as they provide a sequential procedure whose control is inherently represented in the structure.

The use of decision trees to simulate production systems seems to be compatible with a large class of problem solving techniques. For instance, suppose that medical assistance is needed in a remote or inaccessible area. The diagnoses and remedies of any doctor are actually the results of an implicit and intuitive manipulation of a decision-tree-like structure in the doctor's mind. As a result, the doctor's knowledge can be represented

by a decision tree. A computer which has access to this decision tree will approximate a doctor's knowledge and thus can be a good replacement for the doctor in this area. The success or the failure of a doctor to efficiently give his diagnoses and remedies depends on how good this doctor is. A good doctor has to ask only the appropriate questions to his patients and order, accordingly, the appropriate tests to be conducted. Based on all that, a diagnosis and a remedy are prescribed. This process, which is done intuitively by the doctor, is in fact an intelligent manipulation of this medical knowledge base. It is desirable to copy this behavior within the expert system. The expert system must be able to manipulate the decision tree structure using predefined rules. A general mathematical basis for decision processes and, particularly, decision trees has to be established.

Because the knowledge base of expert systems is created by human experts, and because much human knowledge is vague and ill-defined, it is usually true that facts and rules are neither totally certain nor totally consistent. For instance, in our doctor's problem, the situation is not as simple as it seems. The diagnosis process is, at times, heuristic and hypothetical. As a result the decision tree structure, to be constructed, must take into account this important consideration. Several alternatives were proposed to approach the problem (*e.g.*, fuzzy sets [54]). A probabilistic approach to the problem is adopted for this dissertation.

Many problems require the ability to handle uncertain knowledge. In principle the theory of probability provides solutions to these problems. However, the representation and manipulation of large amounts of probabilistic information in a flexible and efficient manner remain a very difficult problem. Such cases arise, for instance, in arranging and manipulating statistical data-bases, probabilistic knowledge-bases, and in large scale state estimation problems. The main issue in these problems is the evaluation of the available representation tools with respect to what one expects when handling uncertain knowledge.

Efficiency and flexibility are naturally a first requirement. The candidate tool must also incorporate probability or, at least, some kind of measure. It is also highly desirable for this representation to be operational even with incomplete information and arbitrary complex statistical dependence. These goals are quite a challenge. The impact of such a representation will affect several areas in computers and engineering. This dissertation attempts to approach these problems by the use of stochastic decision trees, the stochastic version of decision trees.

Decision trees are a much under-rated form of representation. Perhaps one reason for this has been their reputation as being primitive and awkward. When used, decision trees were not always mathematically motivated. As a result, their structure was not fully discovered nor totally exploited. A mathematical foundation for the representation of knowledge using decision trees, however, has been developed by Cockett [16,17,18], who formalized the use of decision trees in the deterministic case, gave an interpretation of these structures in the context of coalgebra theory, and defined a set of complete and sound inference rules which regulate the mechanisms of storage and retrieval involved in the manipulation of such structures. This provides an exact theoretical basis for the development of manipulation algorithms involving decision trees [32,33]. However, as many realistic applications are stochastic in nature, it is just as important to develop the same tools to handle uncertain knowledge. To our knowledge, neither the formal definition of stochastic decision trees nor their structural manipulations has been treated in the literature. Because of a lack of a mathematical foundation, the tools for utilizing the tree structure representation did not exist. The capabilities of manipulating the structure in order to simplify the representation, to update it, or to retrieve information from it were all lacking. Instead, authors used unnecessarily large structures, assumed restraining simplifications, or processed probabilities quite separately from the structure.

This dissertation extends the results of Cockett [16,17,18] pertaining to discrete decision trees to the stochastic case. This extension, in many ways, parallels the developments of discrete decision theory. However, the incorporated probabilities introduce significant differences, which give the theory a rather different flavor. This theory has led to a new approach to the problem of representing and manipulating uncertain knowledge. The approach is closely allied to two well-established representations: Nilsson's semantic trees and Pearl's belief networks. While Nilsson used binary decision trees to conduct the computation of probabilities (outside the structure), Pearl used belief networks—graphs which indicate dependence between the linked propositions. The representation, in this dissertation, uses stochastic terms (graphically represented by stochastic decision trees) which are terms of a *stochastic decision theory*. This representation has the main advantage of having the probabilistic information as an integral part of the representation; stochastic terms can thus be directly manipulated to perform many probability calculations. This differs markedly from the manner in which Nilsson collects the probabilities into matrices to be manipulated quite separately from the tree structure and the manner in which Pearl propagates probabilities without changing the network structure.

Category theory offers a solid mathematical foundation for stochastic decision theory, its terms, and the rules for their structural manipulation. The category of *description spaces* provides a semantic and probabilistic setting in which models for this theory will lie. The main issue in the theory is to show that the manipulations we propose are sound and complete. This is the main concern of Chapter 4.

One of the major problems in uncertain knowledge representation is the problem of incomplete information. This situation almost always arises in real-world applications. For instance, in expert systems, many of the rules provided by the consulted human experts or data collected from the environment are corrupted and not known with certainty. It is

therefore necessary to define a probability distribution on these rules or data to associate a measure of confidence or trust with them. Unfortunately, this probability distribution is not always complete and, in many cases, it is necessary to provide a best estimate, in some sense, of the missing probabilities. The above problem is often referred to as *probabilistic entailment*. The probabilistic entailment problem is addressed in Chapter 5 when the uncertain knowledge is represented by means of a *decision tree* structure.

A solution to the probabilistic entailment problem is proposed and is based upon the *principle of maximum entropy*. *Entropy*, as defined by Shannon [64], is a measure of how uncertain one is about the outcome of a given random experiment. It can also be viewed as a measure of *information* “produced” by this experiment, when performed. If, for a given random experiment, the probability distribution is totally or partially unknown and if an estimate of the missing information is desired, one choice is a probability distribution which assumes minimum a priori information for maximum entropy. This concept is embodied in the *maximum entropy principle*.

Many authors (see for instance [10,13,55,59]) adopted the maximum entropy principle to solve the probabilistic entailment problem. In [10], Cheeseman presented a new method based on the principle of maximum entropy for calculating the most unbiased probability estimate under given constraints. However, similar to all previous approaches, Cheeseman’s method remained computationally complex for realistic problems. The reason for this is perhaps due to the fact that all these approaches used only the joint probability distributions because of the lack of a general capability to systematically manipulate the relationship between the individual conditional probabilities. We significantly alleviate the computational burden by using stochastic decision theory, providing a recursive method to compute entropy. This method significantly simplifies the problem of probabilistic entailment in the context of the maximum entropy principle. Chapter 5 provides a solution

to the problem of incomplete information which often arises in many realistic knowledge representation problems and gives an algorithm for information completion based on the maximum entropy principle.

Many realistic problems in which uncertain knowledge is described by a number of propositions have the *sparse dependence* property. This property means that, in the stochastic decision tree representing the knowledge, many of the propositions (nodes of the tree) are statistically mutually independent from one another. It is then desirable, when dealing with *sparse dependent* structures, to take advantage of this in the representation and in the manipulation of the structure's stochastic terms. The main problem, though, is to efficiently uncover the probabilistic independencies/dependencies among the decisions featured in a given stochastic term. This is the subject of Chapter 6.

The stochastic term manipulations not only change the syntactic structure of the term but also, simultaneously, the probabilistic information attached to it. This fact is perhaps the biggest advantage the stochastic decision theory representation offers in comparison to the existing uncertain knowledge representation tools. As a result, it is possible to correlate basic manipulation rules with familiar probability calculations. This constitutes the primary motivation and contribution of this dissertation. It is not only possible to drive probabilistic calculations by structural manipulations, but also to ascertain whether a particular probability calculation is possible by relating it to a structural requirement, which can be easily checked. The main concern of this dissertation is to use stochastic decision theory to address classical problems in control systems, such as state estimation in a class of finite state systems. More specifically, this dissertation explores the structural properties which affect some familiar probability calculations such as the ones involved in information retrieval or state estimation problems. This is investigated in Chapter 7. This chapter re-casts a state estimation problem in the context of stochastic decision theory and

interprets the estimation process in terms of stochastic term manipulations. The results on probabilistic entailment from Chapter 5 and statistical independence in Chapter 6 are then used to devise a cluster decomposition scheme for stochastic terms which decomposes the problem of state estimation into a number of smaller state estimation problems. This decomposition scheme will significantly alleviate the computational explosion encountered with large scale systems.

The dissertation is organized as follows: Chapter 2 provides a survey of relevant literature. Chapter 3 gives some background concepts used in the dissertation. The first section of this chapter reviews some basic results of estimation theory. The second section introduces the basic concepts of category theory. Cockett's discrete decision theory is introduced in the Section 2.3. The last section reviews some fundamental concepts in measure and probability theory. Chapter 4 defines stochastic decision theory in the mathematical setting of category theory. The first section sets the mathematical foundation of the theory. The second section formally defines stochastic decision theory and develops the basis of stochastic term manipulations. The concept of model satisfaction is discussed in the following section. The last two sections introduce the notions of structural and general retrieval properties of stochastic decision theory and establish the completeness properties of the manipulations. Chapter 5 addresses and solves the problem of probabilistic entailment in stochastic decision theory in the context of the maximum entropy principle. Chapter 6 is a complete characterization of statistical interdependence between decisions in stochastic decision theory. Chapter 7 addresses the problem of state estimation in a class of finite state static systems whose observation spaces are described by stochastic decision trees, and shows how the estimation procedure can be viewed in terms of some

defined structural tree manipulations. A cluster decomposition scheme, for stochastic terms, is then proposed for decomposing the state estimation problem. Chapter 8 is a summary and assessment of the contributions of this research.

CHAPTER 2

RELEVANT LITERATURE

Estimation problems were first addressed by Gauss almost two centuries ago. Gauss was interested in determining orbits of celestial bodies based on many observations. The technique he used has a statistical basis and is known today as the least squares technique. A probabilistic approach to the problem was discovered nearly four decades ago, giving a mathematical interpretation to the least squares solution and widening the scope of the problem. In this dissertation, the estimation problem is addressed from the decision theory point of view in the sense that one makes a decision under uncertainty when delivering an estimate of a given variable. This decision theory setting requires a tool for representing uncertain information or knowledge. This representation tool must also be flexible enough to allow computations similar to the ones involved in estimation problems.

The problem of representing uncertain knowledge has long been and still is a challenge. The impact of such a representation will affect several areas in computers and engineering. Many graphically-inspired tools for uncertain knowledge representation have been proposed, most of which use decision trees and belief networks. Belief networks gained popularity over decision trees even though they may be more expensive in terms of space and efficiency. Decision trees had the reputation of becoming quickly un-readable and un-manageable with increasing size. The reason for this is while belief networks offered a dynamic architecture (to some extent) for knowledge manipulation, decision trees remained a mere illustration tool. The work of Cockett with discrete decision theory may change the situation. Cockett [16,17,18,19,20] has formalized the use of decision trees in

the discrete case. He gave an interpretation of these structures in the context of coalgebra theory and defined a set of complete and sound inference rules which regulates the mechanism of storage and retrieval involved in the manipulation of such structures. This provided an exact theoretical basis for the development of manipulation algorithms involving decision trees [32,33]. However, as many realistic applications are of a stochastic nature, it is just as important to develop the same tools to handle uncertain knowledge.

There have been several attempts to incorporate probabilistic information into knowledge representations. This dissertation extends the results of Cockett pertaining to discrete decision theory to the stochastic case (see in particular [15]) and introduces stochastic decision theory as a flexible and convenient uncertain knowledge representation tool. This chapter reviews some of the literature on uncertain knowledge representation and compares them to the stochastic decision theory representation. The work of Cockett will be thoroughly discussed in Chapter 3.

This chapter briefly reviews some basic results of estimation theory and introduces two of the well-established and accepted tools of uncertain knowledge representation: Nilsson's semantic trees and Pearl's belief networks. We shall briefly review these two tools of knowledge representation and compare them to the stochastic decision theory representation.

2.1 STATE ESTIMATION THEORY

Consider the *finite-dimensional discrete stochastic dynamical* system described by the equation

$$x_{k+1} = \varphi(x_k, t_{k+1}, t_k) + \Gamma(x_k, t_k)\omega_{k+1}, \quad k = 0, 1, \dots, \quad (2.1)$$

where x_k , an n -vector, is the state of the system at t_k , φ an n -vector function, Γ an $n \times r$ -vector function, and $\{\omega_k, k = 1, 2, \dots\}$ an r -vector white Gaussian sequence defined by its mean and covariance ($\omega_k \sim N(0, Q_k)$). The distribution of the initial state x_0 is assumed given, and x_0 is assumed independent of $\{\omega_k\}$. It is well known (see for example [35,26]), in these conditions, that the random sequence $\{x_k\}$ generated by Equation 2.1 is *Markov*. Let the discrete measurement process for the system be described by the equation

$$y_k = h(x_k, t_k) + v_k, \quad k = 1, 2, \dots, \quad (2.2)$$

where y_k , an m -vector, denotes the noisy observation (or measurement) of the state x_k at time t_k , h an m -vector function, and $\{v_k, k = 1, 2, \dots\}$ an m -vector white Gaussian sequence defined by its mean and covariance ($v_k \sim N(0, R_k), R_k > 0$). It is assumed, for simplicity, that $\{\omega_k\}$ and $\{v_k\}$ are independent and $\{\omega_k\}$ is independent of x_0 .

In view of the above, the discrete state estimation problem can be formulated as follows: Given a realization of a sequence of observations from Equation 2.2, say $Y_l = \{y_1, \dots, y_l\}$, compute an estimate of the state x_k based upon Y_l .

There are three possibilities for this problem:

1. $k = l$: This is the *filtering* problem.
2. $k < l$: This is the *smoothing* problem.
3. $k > l$: This is the *prediction* problem.

This dissertation deals only with the discrete filtering problem, although we will refer to it as the state estimation problem throughout.

The linearity of the system dynamics and the observation process constitutes an important special case of the state estimation problem. Linearity means that time and state

are separable in Equations 2.1 and 2.2, *i.e.*,

$$\varphi(x_k, t_{k+1}, t_k) = \Phi(t_{k+1}, t_k) x_k, \quad (2.3)$$

where Φ is an $n \times n$ -matrix, the matrix function Γ is independent of x_k , and

$$h(x_k, t_k) = M(t_k) x_k, \quad (2.4)$$

where M is an $m \times n$ -matrix. The linear state estimation problem admits a particularly tractable solution because the sequence $\{x_k\}$ generated by the dynamics equation (Equation 2.1) is Gaussian as well as Markov. The point is, in this case, the state estimation solution densities are sufficiently described by their means and covariances, while in the non-linear case, the whole functions have to be characterized. In the former case, the filter is said to have a finite-dimensional state (consisting of the mean and the covariance); the filter state is said to be infinite-dimensional in the latter case.

Gauss first addressed the problem of discrete state estimation from a statistical point of view using the *least squares* technique. This approach does not have a probabilistic meaning as the sequences $\{\omega_k\}$ and $\{v_k\}$ of Equations 2.1 and 2.2 are regarded as errors of unknown character rather than random inputs with well-defined statistics. The problem is, given a realization $\{y_1, \dots, y_N\}$ of noisy observations from Equation 2.2, determine estimates for $\{x_0, x_1, \dots, x_N\}$ so that, simultaneously, the errors in the dynamical system and the observation process are minimized. This minimization is done with respect to the sequence $\{x_0, x_1, \dots, x_N; \omega_1, \dots, \omega_N\}$ of the following *cost function*:

$$\begin{aligned} J_N = & \frac{1}{2}(x_0 - \bar{x}_0)^T P_0^{-1}(x_0 - \bar{x}_0) + \frac{1}{2} \sum_{k=1}^N [y_k - h(x_k, t_k)]^T R_k^{-1} [y_k - h(x_k, t_k)] \\ & + \frac{1}{2} \sum_{k=1}^N \omega_k^T Q_k^{-1} \omega_k, \end{aligned} \quad (2.5)$$

subject to the constraints of Equation 2.1.

The first term in J_N reflects the belief in the prior estimate $\bar{x}_0$ of x_0 , the second term reflects our belief in the measurement model, while the last term reflects the belief in the dynamics of the system. P_0^{-1} , R_k^{-1} , and Q_k^{-1} are the positive definite matrices which quantify these beliefs. In practice, these matrices play the role of adjustment parameters for the best least squares fit.

The re-interpretation of the sequences $\{\omega_k\}$ and $\{v_k\}$ of Equations 2.1 and 2.2 as random sequences constitutes the so-called probabilistic approach to state estimation. In their pioneering work, Kalman and Bucy [39] and Kalman [38] adopted the probabilistic approach and completely solved this problem in the linear case. The non-linear case was first treated by Kushner [42,43] and Stratonovich [66,67]. The probabilistic approach adopts the so-called Bayesian point of view which considers the conditional probability density function

$$p(x_k | Y_k) \tag{2.6}$$

as the complete solution to the discrete filtering problem. Once Equation 2.6 is obtained, there still be the question of which estimate of the state should be selected. A criterion of closeness to the actual state must be set, and the estimate which minimizes it is optimal and must be selected. In [35], Jazwinski showed that for a class of loss criteria and when the conditional probability density function of Equation 2.6 is symmetric about its mean and unimodal, this conditional mean is the optimal choice. These conditions are clearly met when the conditional probability density function is Gaussian; this then constitutes the solution to the linear state estimation problem. A particularly important solution for this problem is a solution obtained for a quadratic loss function; this solution is called the *minimum variance* or the *minimum mean square error* solution of the state estimation problem. It turned out that the least squares problem defined in Equation 2.5 is equivalent

to maximizing the conditional probability density function

$$p(x_0, \dots, x_N \mid Y_N) \quad (2.7)$$

with respect to $\{x_0, \dots, x_N\}$ provided that the function Γ in Equation 2.1 is independent of x_k . This is the *joint maximum likelihood (Bayesian) estimation*. The solution is the *peak* or *mode* of the conditional probability density function of Equation 2.7. The *maximum likelihood estimate* obtained by maximizing the marginal conditional probability density function

$$p(x_N \mid Y_N) \quad (2.8)$$

is, in general, different from joint maximum likelihood estimate. The two estimates coincide in the linear case.

The statistical approach (without a probabilistic interpretation) to the state estimation problem is conceptually simpler than the probabilistic approach. There is no need to make assumptions about the sequences $\{\omega_k\}$ and $\{v_k\}$; they are viewed as deterministic errors with arbitrary character. On the other hand, it is difficult to interpret the results obtained with the statistical approach. The probabilistic approach is conceptually harder, but provides a solid mathematical interpretation for the solution. Both methods, however, suffer, when implemented, from computational and storage problems. Many authors proposed suboptimal filters to avoid this problem, *e.g.*, the *batch processing* filter and the *recursive least squares* filter (see [35] for more details).

The stochastic finite-dimensional case discussed above includes an important large class of systems which can be described by maps between finite sets. Such a system is referred to as a *stochastic sequential machine* or a *stochastic finite automaton* [6,23,45,57] and is formally defined as a quadruple

$$M = (S, X, Y, \{A(Y/X)\}), \quad (2.9)$$

where S, X , and Y are finite sets of internal states, inputs, and outputs, respectively, and $\{A(Y/X)\}$ is a finite set containing $\text{card}(X) \times \text{card}(Y)$ square matrices of order $\text{card}(S)$. A matrix $[a_{ij}(y/x)]$ with $i, j \in S$, $y \in Y$, and $x \in X$, has for elements $a_{ij}(y/x)$, the probability of the machine going to state j and generating output y , given it had been in state i and was given input x . Defining the vector

$$\begin{cases} \eta(y/x) &= A(y/x)\eta \\ \eta(\lambda/\lambda) &= \eta \end{cases} \quad (2.10)$$

with η denoting a column vector with all entries equal to 1 and λ the empty input, the i^{th} entry of the vector $\eta(y/x)$ represents the probability of the machine generating output y (and moving to some state) having been in state i and given input x . The information about the current state of a machine can be completely determined by the so-called *information vector*

$$\Pi = [\pi_1, \pi_2, \dots, \pi_n], \quad (2.11)$$

with π_i being the probability that the current internal state of the machine is i . Defining

$$\begin{cases} \Pi(y/x) &= \Pi A(y/x) \\ \Pi(\lambda/\lambda) &= \Pi \\ p_{\Pi}(y/x) &= \Pi \eta(y/x) \\ &= \Pi(y/x)\eta, \end{cases} \quad (2.12)$$

it follows that $p_{\Pi}(y/x)$ is the probability of the machine generating output y when started with the distribution Π over its states and given input x . Also, the j^{th} entry of the vector $\Pi(y/x)$ represents the probability of the machine generating output y and moving to state j when started with the distribution Π over its states and given input x .

A transition-to-transition update of the information vector can be constructed as follows: If Π is the current information vector, the input has been x , and the output is y ,

then the new information vector $\bar{\Pi}(y/x)$ is given by

$$\bar{\pi}_i(y/x) = \begin{cases} \frac{\bar{\Pi}(y/x)}{p_{\Pi}(y/x)} & \text{if } p_{\Pi}(y/x) \neq 0 \\ \text{undefined} & \text{otherwise} \end{cases} \quad (2.13)$$

This formulation allows us to treat a sequence of inputs or outputs as a single input or output. This enables the lumping of transitions into a single one. For the matrix A , a sequence of inputs/outputs can be decomposed as follows:

$$A(v/u) = [a_{ij}(v/u)] = A(y_1/x_1)A(y_2/x_2) \cdots A(y_k/x_k) \quad (2.14)$$

$$a_{ij}(vy/ux) = \sum_k a_{ik}(v/u)a_{kj}(y/x) \quad (2.15)$$

$$\eta(vy/ux) = A(v/u)\eta(y/x) \quad (2.16)$$

$$p_{\Pi}(vy/ux) = p_{\Pi}(v/u)p_{\Pi(v,u)}(y/x) \quad (2.17)$$

Finite state automata are assumed to have the important property (the Markov property) that every transition (from a current state to a new state) is only dependent upon the current state and input and is completely independent from the previous states and inputs of the machine. When, in addition, these transitions are independent from the inputs of the machine, in which case, the machine consists of a finite number of states superseding one another at some discrete times, the machine is referred to as a *discrete Markov chain*. For a complete investigation of the properties of Markov chains, see [62] and [61].

The one-step Bayesian solution to the state estimation problem as formulated in the finite-dimensional case is provided, in the finite state case, by Equation 2.13. The recursive solution for the problem is given by Equation 2.17 and can be found in [7].

2.2 NILSSON'S SEMANTIC TREES

Nilsson [55] proposed *binary semantic trees* for uncertain knowledge representation. This representation is a semantic generalization of logic in which the truth value of a *sentence* or *decision* is taken to be the probability of occurrence of that sentence. Each one of these sentences can be either *true* or *false*. These binary semantic trees are the graphical representation of the truth variables of the sentences in which the inconsistent paths are simply closed off. For a particular sentence, there can be two sets of possible *worlds*, one containing worlds in which the sentence is true and the other containing worlds in which the sentence is false. The uncertainty about the actual world is given by the probability $p \in [0, 1]$ of the sentence being true. If we have more sentences, we can define more sets of possible worlds. Each set of worlds contains a unique and consistent set of truth values for the sentences. If we have L sentences, we might have as many as 2^L sets of possible worlds. Typically, though, we will have fewer worlds as some of the truth value combinations might be logically inconsistent. The following example (given in [55]) illustrates the representation.

Example 1 Consider the binary semantic tree of Figure 2.1 depicting the logical sentences P , $P \supset Q$, and Q in which the inconsistent paths of the tree are simply closed off. To process probabilities, the author introduced the vectors $\mathbf{P}$ and $\mathbf{\Pi}$ and the matrix $\mathbf{V}$. In this example, the vector $\mathbf{P}$ is 4-dimensional column vector whose i^{th} component p_i represents the probability of the i^{th} set of possible worlds. The set of possible worlds is characterized

by the different consistent truth valuations that can be given to the four sentences, namely,

$$\mathbf{V}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{V}_2 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{V}_3 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{V}_4 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

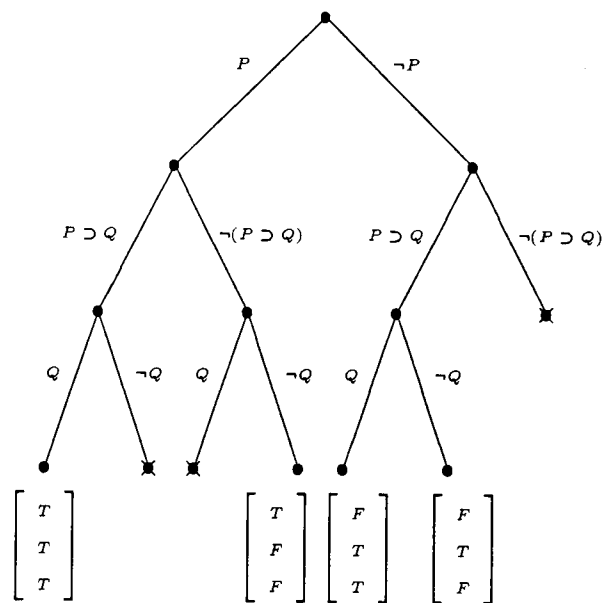


Figure 2.1: Nilsson's Representation.

Letting the matrix $\mathbf{V} = [\mathbf{V}_1 \mid \dots \mid \mathbf{V}_4]$, if the i^{th} component π_i of the 3-dimensional column vector $\mathbf{\Pi}$ denotes the probability of the i^{th} sentence, the probability of the sentences can then be related to the probability of the possible worlds by $\mathbf{\Pi} = \mathbf{VP}$, or

$$\begin{bmatrix} \pi_1 = p_1 + p_2 \\ \pi_2 = p_1 + p_3 + p_4 \\ \pi_3 = p_1 + p_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix}.$$

□

Nilsson addressed the problem of probabilistic entailment of sentences given the joint probability distribution and gave a graphical interpretation of the region of consistent probability assignment in the probability space. He gave an approximate solution to the problem based on the method of *projection approximation* and showed that the solution coincided with the one obtained according to the *maximum entropy principle*.

Nilsson also addressed the problem of updating the network representation when new information is received by means of Bayes' rule of probability update. The author did not elaborate thoroughly on the subject, as his structure lacked manipulation capabilities.

Few remarks can be made about Nilsson's approach. Nilsson did not incorporate the probability distribution into the tree structure and, as a result, probability calculations had to be carried out separately from the structure. Furthermore, the matrix form he used for these calculations is restricted to binary decisions and to trees with complete information and is not readily extendable to the non-binary case. Nilsson's semantic trees are primarily used for illustration purposes as no manipulation, at the tree level, is attempted.

2.3 PEARL'S BELIEF NETWORKS

The work of Pearl [58] on *belief networks* was motivated by his attempt to model human's inferential reasoning. Belief networks can also be used to represent generic knowledge of a specific-domain expert. These networks are directed acyclic graphs in which the nodes represent propositions (or variables), the arcs indicate direct dependencies between the linked propositions. Pearl gave a method for constructing belief networks based on the chain rule representation of joint probability distribution and a judgment on the dependence (independence) relationship among the knowledge propositions. Pearl also outlined

a graphical method establishing a correspondence between the topology of the network and familiar independence concepts.

Example 2 (The Murder Case)

Consider the example given in Pearl [58] of three murder suspects, one of whom has definitely committed the crime. The murder weapon, with fingerprints, was later found by the police. Let A stand for the identity of the killer. Let B stand for the identity of the person whose fingerprints were left on the weapon, and let C represent the possible readings of the fingerprint testing. The relations between these three decisions would normally be conceptualized by the following belief network

$$A \longrightarrow B \longrightarrow C.$$

Each one of the decisions has three branches or *siblings* corresponding to the three possible identities. Suppose that, under normal circumstances, the killer is likely to leave his fingerprint on the weapon and thus the following probability distribution is to be used in the trial:

$$p(B_j | A_i) = \begin{cases} 0.80, & \text{if } A_i = B_j, i, j = 1, 2, 3, \\ 0.10, & \text{if } A_i \neq B_j, i, j = 1, 2, 3. \end{cases}$$

Perhaps the merit of belief networks resides in the fact that this structure is not only used to merely store knowledge data, but also to direct and activate the data flow computations which manipulate the knowledge. Pearl [58] outlined an algorithm which updates the network probability distribution when new information or evidence is received. Evidence can be either *definite* or *virtual* [58]. Definite evidence corresponds to direct observations which validate with certainty the values of some decisions in the structure. Virtual evidence corresponds to judgments based on undisclosed observations which affect the belief in some decision in the structure. The new information or evidence

is modeled by an additional node representing the undisclosed observation and through which the information is injected through the network. The algorithm traces the flow of this information throughout the network decisions until an equilibrium (or a steady state) is reached.

This algorithm, however, is only possible with singly connected networks and is quite complex even for small-size structures.

Example 3 (The Murder Case (Cont'))

Continuing with the example of the murder case, let D_B^- represent the results of the fingerprint report and D_B^+ all the other information related to the trial. The quantity $p(B_i | D_B^+)$ will then represents our belief, prior to the fingerprint report, that the i^{th} man was last to hold the weapon. The quantity $p(D_B^- | B_i)$ is the fingerprint report and is, as before, $p(D_B^- | B_i) = (0.80, 0.60, 0.50)$. Assuming that $p(B | D_B^+) = (0.60, 0.30, 0.10)$, the probability of each suspect being the killer is given by, according to Pearl, by

$$\begin{aligned} p(B) &= \alpha p(D_B^- | B) p(B | D_B^+) \\ &= \alpha (0.80, 0.60, 0.50) (0.60, 0.30, 0.10) \\ &= \alpha (0.48, 0.18, 0.05) \\ &= (0.676, 0.254, 0.07), \end{aligned}$$

where, in the last equation, a proper normalization factor was introduced.

The following remarks can be made about the belief network representation. The method assumes complete information and total knowledge of the inter-dependence between the propositions (nodes) of the network. The graphical method given by Pearl which uncovers, in a given belief network, all the (conditional) independence between the network propositions. Although this information can be quite useful, it is not clear how it

can be implemented efficiently. The problem is, for a given belief network, not every independence relation is transparent. In fact, belief networks are strongly dependent upon the selected order of the network propositions, meaning that a different network is obtained for each possible ordering of these propositions. This is a major problem as the author derived different treatments for different kind of networks.

The belief network structure will be, in general, smaller in size than the tree structure, however the storage requirements can be much higher than in the tree structure. Furthermore, belief network structures are strongly dependent upon the arbitrarily chosen decision order, and probability calculations can be quite complex even for small-size problems and not always possible. For instance, only *tree-structured* belief networks (networks in which every node, except the root, has only one incoming node, meaning that it depends upon only one decision) and *singly connected* belief networks (networks in which a node is allowed to have multiple parents, but the graph is required to be singly connected, namely, there could be only one undirected path, at most, between any two nodes) can be processed by Pearl's algorithm. Networks with arbitrary inter-dependence liaisons between decisions (*multiply connected* networks) may be transformed, under certain conditions, into tree-structured networks; otherwise no processing algorithm is given. This transformation, when it is possible, introduces some auxiliary (dummy) variables in the representation. This extension, however, has the major drawback of considerably augmenting the size of the network.

2.4 THE STOCHASTIC DECISION THEORY REPRESENTATION

The stochastic decision theory approach to uncertain knowledge representation is somewhat similar to both of Nilsson's and Pearl's representations. Like Nilsson's representa-

tion, it uses decision trees, deals with inconsistencies, and tolerates incompleteness. Like Pearl's representation, it also provides a dynamical structure for information updates and retrievals. Unlike any other representation, the stochastic decision theory approach actually unifies the structure with its probabilistic data and provides structural manipulation capabilities which also incorporate the probabilistic information.

The following example shows how Nilsson's semantic trees can be modified into the stochastic decision theory version of the representation. First, the probabilistic information is incorporated into the tree structure, and second, the inconsistent paths are eliminated rather than closed off.

Example 4 Reconsider Nilsson's Example 1. Figure 2.2 shows how the stochastic decision theory representation is constructed from Nilsson's semantic tree of that example. The leaves of the tree are first labeled by the probabilities of occurrence of the corresponding paths. These probabilities correspond, in Nilsson's representation, to the probability vector $\mathbf{P}$. Summing over the appropriate leaves' probabilities, the conditional probabilities of the sentences are easily computed. To obtain the unconditioned probabilities derived by Nilsson, the transposition manipulation (Chapter 4, on Page 76) may be used to put the desired decision at the top of the tree; this decision will then be labeled by the probabilities obtained by Nilsson. For instance, since the decision P is already at the top of the tree, this decision is labeled by the probability $p_1 + p_2$, which is the first component π_1 of $\mathbf{\Pi}$ in Nilsson's representation.

In fact, the stochastic decision theory representation does not represent inconsistent paths; rather, it eliminates them. Figure 2.3 gives the stochastic decision theory representation for the example.

□

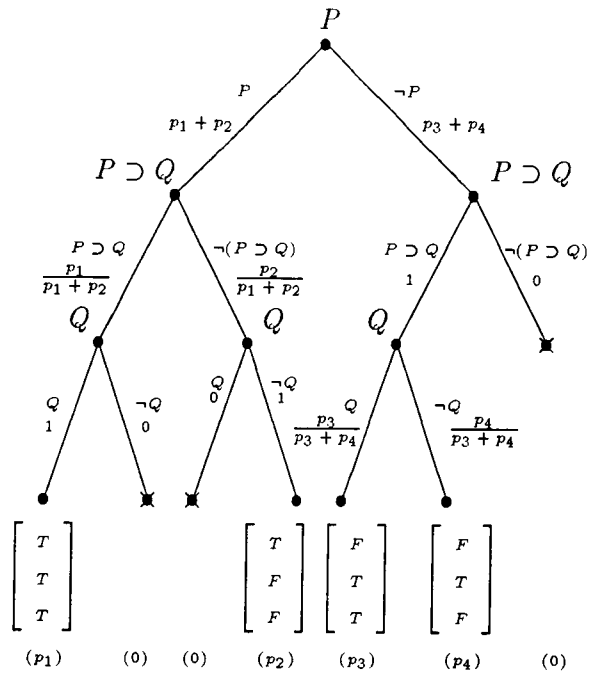


Figure 2.2: Stochastic Decision Theory Construction.

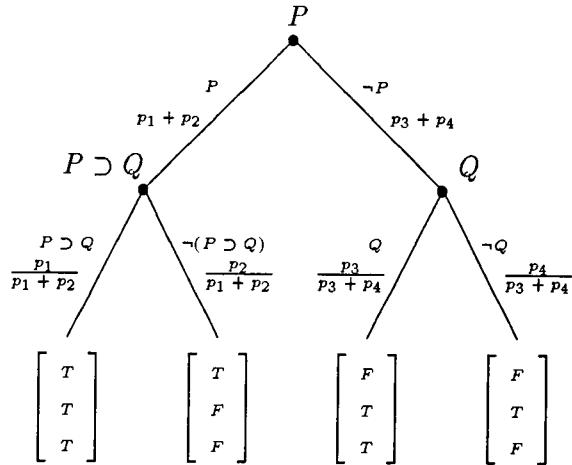


Figure 2.3: Stochastic Decision Theory Representation.

As to the belief network representation, the stochastic decision theory representation has to be constructed from the probabilistic data as the two structures are conceptually different. The following example illustrates the procedure.

Example 5 (The Murder Case (Cont'))

From the information provided in Pearl's example, one can determine, uniquely, the initial belief in the assertion $A = A_i$ which is computed by transposing the decision B to the top of the tree. This gives $p(A) = (\frac{5}{7}, \frac{2}{7}, 0)$. The stochastic decision term representing this example is depicted in Figure 2.4 and is bigger in size than the corresponding belief network, because it explicitly depicts the branches of the decisions.

□

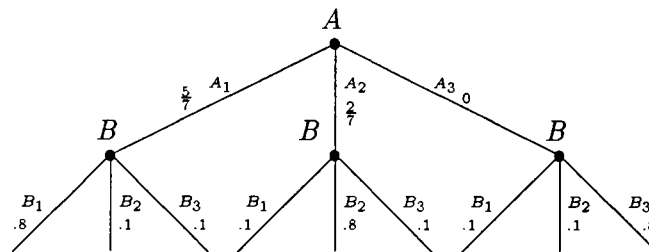


Figure 2.4: Stochastic Decision Tree Representation.

The problem, addressed by Pearl and Nilsson, of updating network representations when new information is received, is closely related to a well known classical problem in control systems, namely the *state estimation* problem. The control system in question is supposed to be a *finite state* system whose initial probability distribution is represented by a stochastic decision tree structure. The observation process is incorporated into the

structure of the tree describing the system by viewing each incoming observation as an additional decision, called the *observation decision*, which is appended to the leaves of the original tree, similarly to Nilsson's and Pearl's updating procedure. It will be shown in Chapter 7 that the estimation procedure is equivalent to transposing the observation decision to the top of the tree.

Example 6 (The Murder Trial (Cont'))

In Example 2, in order to model the reliability of the fingerprint laboratory test, Pearl used a matrix $p(C_k | B_j)$, satisfying $\sum_k p(C_k | B_j) = 1, \forall j$. Each entry in this matrix is an *if-then* rule of the type:

If the fingerprint is of suspect B_j , then
 expect reading C_k , with certainty $p(C_k | B_j)$.

Suppose that the fingerprint report is $p(C_{observed} | B) = (0.80, 0.60, 0.50)$, meaning that there is an 80% chance the fingerprints left on the murder weapon belong to suspect B_1 , 60% chance it belong to suspect B_2 , and 50% chance it belong to suspect B_3 .

In this example Pearl assumed a *a priori* probability of $p(B) = (0.60, 0.30, 0.10)$ which gives the belief in the identity of the last weapon holder based upon all the other trial testimonies, but before the fingerprint report. Our purpose is then to update this belief after the fingerprint report. This problem corresponds, in our stochastic decision theory representation, to first transposing the decision B to the top of the tree in Figure 2.4 to obtain the decision tree of Figure 2.5, and second to appending to that tree the observation decision, C , from the laboratory report. This leads to the decision tree of Figure 2.6.

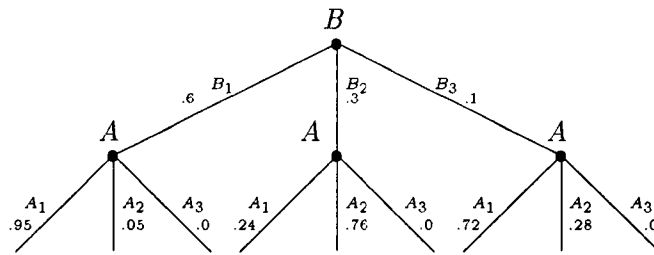


Figure 2.5: Transposed Stochastic Decision Tree.

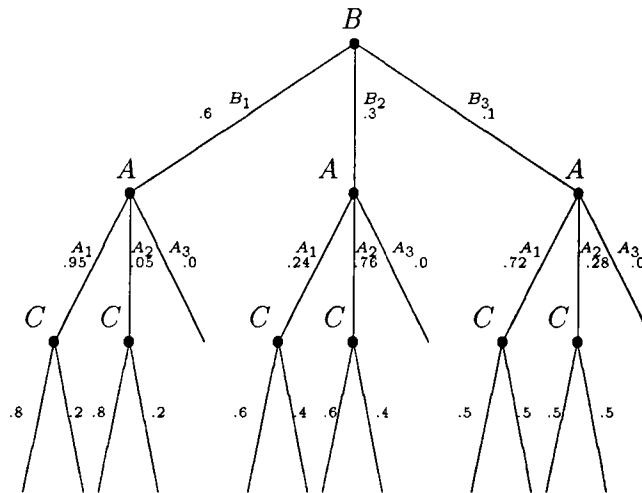


Figure 2.6: The Augmented Decision Tree.

Moving the observation decision C to the top of the tree, the stochastic decision tree of Figure 2.7 is obtained. The subtree rooted at B in Figure 2.7 is labeled with the desired updated probabilities which agree with Pearl's results, namely, $p(B \mid C_{\text{observed}}) = (0.64, 0.26, 0.1)$. Notice that as the observation decision (the fingerprint report) is independent of the decision A (the killer identity), we could have appended the observation

decision C below the decision B (the last holder identity). In this way, we avoid the first step in the transposition manipulation in which C is moved above the decision A .

□

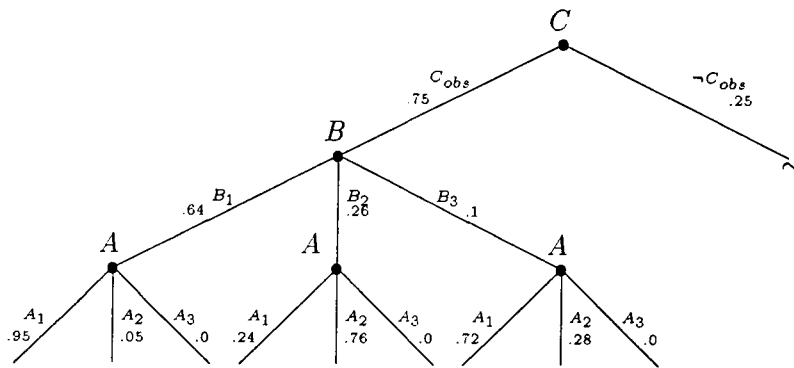


Figure 2.7: The Updated Decision Tree.

CHAPTER 3

BACKGROUND

Any decision-making process is an intelligent manipulation of some acquiescent domain-specific knowledge (combination of facts and rules) for the purpose of producing new facts or rules about the domain. The mathematical theory of categories can be a good vehicle for this decision process. Discrete decision theory uses category theoretic concepts to model decision processes. In order to deal with uncertainty, we intend, in this dissertation, to adopt a parallel approach to discrete decision theory in order to set similar foundations to stochastic decision theory. Measure and probability theories are the main tools in this approach.

This chapter briefly introduces the basic concepts used in this dissertation. Section 3.1 introduces the fundamental concepts of category theory along with some examples and shows how this theory can be useful in modeling decision processes. Section 3.2 briefly presents the work of Cockett [15,16,17] in which he formally defined discrete decision theory in the context of coalgebras. Section 3.3 gives the tools from measure and probability theory used in the definition of stochastic decision theory.

3.1 CATEGORY THEORY

Category theory, a relatively new branch of mathematics, provides a broad abstract perspective for many already well-known mathematical entities. Although many have often labeled it “abstract nonsense” and questioned its usefulness, others demonstrated

how the theory can be successfully and eloquently applied to practical problems in control systems and computer science (see for instance [49]). The main motivation of this theory is to offer an alternative to the classical set membership relation as the main tool for developing mathematical constructions and expressing structural properties. This new alternative relies on the concept of *morphisms*, abstracted from functions or mappings. For instance, in order to define the properties of a collection, rather than consider the collection's members, the link of this collection with other collections will determine the collection's properties.

In this section, we intend to present, as briefly as possible, the theory of categories and show how it can be used to model decision processes. The curious reader is referred to any of the notable books on the subject [50,47,36,27,40].

Definition 1 *A directed graph consists of objects $A, B, C, \dots$, morphisms $f, g, h, \dots$, and two operations:*

- **Domain**, which assigns to each morphism f an object $A = \text{dom } f$;
- **Codomain**, which assigns to each morphism f an object $B = \text{codom } f$.

These operations on f are best indicated by representing the morphism f as an arrow starting at its domain and ending at its codomain as $f : A \longrightarrow B$.

Definition 2 *A category is a directed graph together with two additional operations:*

- **Identity**, which assigns to each object A a morphism $\text{id}_A = 1_A : A \longrightarrow A$;
- **Composition**, which assigns to each pair $\langle g, f \rangle$ of morphisms with $\text{dom } g = \text{codom } f$ a morphism fg^1 , called their composite, with $fg : \text{dom } f \longrightarrow \text{codom } g$.

¹Throughout this dissertation, the composition to the right notation is used.

These two operations are required, in a category, to obey the associativity axiom and the unit law axiom. The former axiom is the usual property of associativity stating that for given objects and morphisms in the following configuration:

$$A \xrightarrow{f} B \xrightarrow{g} C \xrightarrow{h} D,$$

one has the equality

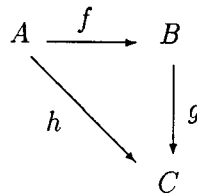
$$f(gh) = (fg)h.$$

The latter axiom states that the identity morphism 1_B of any object B is the identity for the composition operation. This means that for any object B , we have that

$$f1_B = f \quad \text{and} \quad 1_B g = g.$$

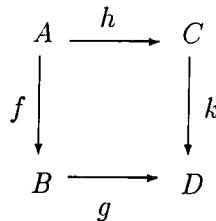
These identities in categories can also be pictured by diagrams consisting of vertices, labeled by the objects of the category, and edges, labeled by the morphisms.

Definition 3 *The following diagram*



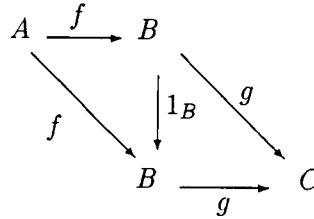
*is called a **commutative** diagram if $fg = h$.*

*We shall say in this case that the morphism h **factors through** B . Likewise, a diagram of the form*



is commutative if $fg = hk$.

For example, the axiom of the unit law simply means that for all objects A, B , and C and for all morphisms f and g , the following diagram is commutative



Categorical duality is the process “Reverse all arrows”. The dual of any statement Σ about a given category is the statement Σ^* formed by making the following replacements throughout in Σ : “domain” by “codomain”, “codomain” by “domain”, and “ h is the composite of g with f ” by “ h is the composite of f with g ”; arrows and composites are reversed. Logic (and, or, ...) are unchanged. The dual category of a category is defined as follows:

Definition 4 *The dual category (or opposite category) of a category $\mathcal{A}$, denoted by $\mathcal{A}^*$, is a category whose objects are the objects of the category $\mathcal{A}$ and whose morphisms are in one-to-one correspondence with the morphisms of $\mathcal{A}$. For each morphism $f : A \longrightarrow B$ of $\mathcal{A}$ corresponds a morphism in $\mathcal{A}^*$, $f^* : B \longrightarrow A$ (the direction is reversed). The composite, in $\mathcal{A}^*$, of two morphisms f^* and g^* is defined by $g^*f^* = (fg)^*$.*

It can be shown that the construction described above defines a category. Any result, proved true for a given category, corresponds to a result, its dual, in the dual category.

A morphism $f : A \longrightarrow B$ is called a *coretraction* if there is a morphism $f' : B \longrightarrow A$ such that $ff' = 1_A$. Dually we say that f is a *retraction* if there is a morphism $f'' : B \longrightarrow A$ such that $f''f = 1_B$. If f is both a retraction and a coretraction, it is called

an *isomorphism*, in which case $f' = f'' = f^{-1}$, the *inverse* of f . A morphism whose codomain is the same as its domain is called an *endomorphism*. An endomorphism which is also an isomorphism is called an *automorphism*. A morphism $f : A \longrightarrow B$ is called a *monomorphism*, or is said to be *monic*, if whenever $\alpha f = \beta f$, we have that $\alpha = \beta$, for all pairs of morphisms (α, β) with codomain A . A morphism f is called an *epimorphism* if $f\alpha = f\beta$ implies that $\alpha = \beta$.

Examples of categories are numerous. The category **Set** has as objects all sets and as morphisms all functions between sets, with the usual identity function and the usual composition of functions. The category **Grp** has groups as objects and group homomorphisms as morphisms. The Category **Top** is formed by all topological spaces together with continuous functions between topological spaces. Vector spaces and linear transformations form the category **Vect**. The category **Pos** is formed by partially ordered sets and monotone functions, and the list goes on. In the context of decision theory, the categories **Set** and **Pos** are most useful.

Definition 5 Given two morphisms $\alpha_1 : A_1 \longrightarrow A$ and $\alpha_2 : A_2 \longrightarrow A$ with a common codomain, a commutative diagram

$$\begin{array}{ccc} P & \xrightarrow{\beta_2} & A_2 \\ \beta_1 \downarrow & & \downarrow \alpha_2 \\ A_1 & \xrightarrow{\alpha_1} & A \end{array}$$

is called a **pullback** for α_1 and α_2 if for every pair of morphisms $\beta'_1 : P' \longrightarrow A_1$ and $\beta'_2 : P' \longrightarrow A_2$ such that $\beta'_1 \alpha_1 = \beta'_2 \alpha_2$, there exists a unique morphism $\gamma : P' \longrightarrow P$ such that $\beta'_1 = \gamma \beta_1$ and $\beta'_2 = \gamma \beta_2$. The object P , referred to as the pullback of α_1 and α_2 , is denoted $\alpha_1 \mid \alpha_2$. The maps β_1 and β_2 are called the projection maps and denoted by

$p(1, \alpha_1 \mid \alpha_2)$ and $p(2, \alpha_1 \mid \alpha_2)$, respectively.

The dual of a pullback is a *pushout*. Thus, a pushout diagram is obtained by reversing the direction of all the arrows in the pullback diagram.

Definition 6 *An object F is called a **final object** for a category $\mathcal{A}$ if for every object A in the category there exists a unique morphism $A \longrightarrow F$. In the dual category, the final object becomes the **initial object**.*

The properties of categories with final objects and in the presence of other assumptions are of importance in the setting of decision theory. These categories uniquely tie attribute maps to decisions (see [18]).

Definition 7 *Let $\{A_i\}_{i \in I}$ be a set of objects in a category $\mathcal{A}$. A **product** [50] for the family is an object A for which any family of morphisms $\{p_i : A \longrightarrow A_i\}_{i \in I}$, called **projections**, with the property that for any family of morphisms $\{\alpha_i : A' \longrightarrow A_i\}_{i \in I}$ there is a unique morphism $\alpha : A' \longrightarrow A$ such that $\alpha p_i = \alpha_i$ for all $i \in I$. The object A will be denoted by $\otimes_{i \in I} A_i$. For a finite family of objects $A_1, A_2, \dots, A_n$ the product is denoted as $A_1 \otimes A_2 \cdots \otimes A_n$.*

Definition 8 *The **coproduct** [50] of the family $\{A_i\}_{i \in I}$ is defined dually to the product. The coproduct is thus an object A for which any family of morphisms $\{u_i : A_i \longrightarrow A\}_{i \in I}$, called **injections**, with the property that for any family of morphisms $\{\alpha_i : A_i \longrightarrow A'\}_{i \in I}$ there is a unique morphism $\alpha : A \longrightarrow A'$ with $u_i \alpha = \alpha_i$ for all $i \in I$. The object A will be denoted as $\oplus_{i \in I} A_i$. For a finite family of objects $A_1, A_2, \dots, A_n$ the coproduct is denoted as $A_1 \oplus A_2 \cdots \oplus A_n$.*

The concept of coproducts turned out to be very important in modeling decision processes. Coproduct objects have the effect of partitioning their base object. This partition,

when disjoint, happens to be the effect one expects from a decision. Disjoint coproducts, as Cockett notes in [18], are the main ingredient in *open* categories. As a result, open categories will be our focus in this dissertation.

For the sake of clarity, we denote, in this dissertation, the injections u_i defined in the coproduct object by b_i^A . These injections are sometimes referred to as *inclusion maps*. Also, the unique map α is denoted, in the product definition, by $\langle \alpha_1, \alpha_2, \dots, \alpha_n \rangle$ and, in the coproduct definition, by $\langle \alpha_1, \alpha_2, \dots, \alpha_n \rangle$ to emphasize duality. The following diagram shows the product object $A = A_1 \otimes A_2$ of two objects A_1 and A_2 :

$$\begin{array}{ccc}
 A_1 \otimes A_2 & \xrightarrow{p_1} & A_1 \\
 \downarrow p_2 & \searrow \alpha & \uparrow \alpha_1 \\
 A_2 & \xleftarrow{\alpha_2} & A'
 \end{array}$$

The following is a diagram of the coproduct object of A_1 and A_2 :

$$\begin{array}{ccc}
 A_1 \oplus A_2 & \xleftarrow{u_1 = b_1^{A_1 \oplus A_2}} & A_1 \\
 \uparrow u_2 = b_2^{A_1 \oplus A_2} & \searrow \alpha & \downarrow \alpha_1 \\
 A_2 & \xrightarrow{\alpha_2} & A'
 \end{array}$$

Definition 9 A product object for the family of objects $\{A_i\}_{i \in I}$ in which $A_i = A$ for any i is denoted by A^I and is referred to as the **n-fold product** [50] of the object A . The coprod-

uct of the same family is denoted by ${}^I A$ and is referred to as the **n-fold coproduct** [50] of the object A . The **diagonal morphism** Δ , with reference to the product/coproduct example diagram, is defined as the map

$$\Delta : A \longrightarrow A^I \quad \text{such that } \Delta p_i = 1_A \quad \forall i \in I.$$

Dually, the **codiagonal morphism** ∇ is defined as the map

$$\nabla : {}^I A \longrightarrow A \quad \text{such that } u_i \nabla = 1_A \quad \forall i \in I.$$

Definition 10 An **operation** [16] in a category $\mathcal{A}$ is a map from an n -fold product of a base object back to the base object itself: $w : Y \otimes Y \otimes \cdots \otimes Y \longrightarrow Y$.

Definition 11 An **algebra** [16] consists of an object together with a set of operations on that object.

The dual concept [50] of these definitions is more useful for decision processes.

Definition 12 A **cooperation** [16] w , on an object X , is a map from X to the **n-fold coproduct** [50] (disjoint union) of X , $w : X \longrightarrow X \oplus X \oplus \cdots \oplus X$. The **arity** of w is the number of copies of X present in the coproduct.

These are the maps which will produce the decision maps as it will be later shown.

Definition 13 A **coalgebra** is an object together with a set of cooperations on that object.

Definition 14 The **trace** of a cooperation w on X is the endomorphism $tr(w)$, defined by the following commutative diagram, where ∇ is the codiagonal morphism of the coproduct as defined above,

$$\begin{array}{ccc}
X & \xrightarrow{w} & X \oplus \cdots \oplus X \\
& \searrow \text{tr}(w) & \downarrow \nabla = \langle 1_X, \dots, 1_X \rangle \\
& & X
\end{array}$$

Note that the codiagonal morphism ∇ can also be represented by $\langle 1_X, 1_X, \dots, 1_X \rangle$.

Definition 15 A cooperation w is said to be **idempotent** if $\text{tr}(w) = \text{id}(X)$. An idempotent cooperation is called a **decision**.

Idempotence is a key property for a cooperation to be a decision. In fact, the repetition and the distribution laws, which we will introduce in the upcoming section, are simple consequences of idempotence in the discrete decision theory setting.

Definition 16 A category $\mathcal{A}$ has **disjoint coproducts** if

- (i) $\mathcal{A}$ has a strict initial object, i.e., any map to the initial object is an isomorphism.
- (ii) Coproducts are stable, i.e., whenever the pullbacks $f \mid g_1$ and $f \mid g_2$ exist, the diagram

$$\begin{array}{ccc}
A & \xrightarrow{f} & B \\
\uparrow \langle p(1, f \mid g_1), p(1, f \mid g_2) \rangle & & \uparrow \langle g_1, g_2 \rangle \\
f \mid g_1 \oplus f \mid g_2 & \xrightarrow{\quad} & C_1 \oplus C_2 \\
p(2, f \mid g_1) \oplus p(2, f \mid g_2) & &
\end{array}$$

is a pullback for all maps $f : A \longrightarrow B$, $g_1 : C_1 \longrightarrow B$, and $g_2 : C_2 \longrightarrow B$.

- (iii) The inclusion maps of the coproduct object $C_1 \oplus C_2$, $b_1^{C_1 \oplus C_2} : C_1 \longrightarrow C_1 \oplus C_2$ and $b_2^{C_1 \oplus C_2} : C_2 \longrightarrow C_1 \oplus C_2$ are monic.
- (iv) The object $b_1^{C_1 \oplus C_2} \mid b_2^{C_1 \oplus C_2}$ is the initial object for A_1 and A_2 .

It is well known that categories of sets, and indeed any topos [36], have disjoint coproducts. In a decision theory settings, it is necessary that the coproducts exist in the category and that they be disjoint. Furthermore, the embedding maps for the coproduct must belong to a set of monics for which pullbacks always exist. Such a setting arises in categories which have image factorization systems.

Definition 17 (Image Factorization System)

For a category $\mathcal{A}$, an **image factorization system** is a pair (E, M) , where E and M are classes of morphisms in $\mathcal{A}$ satisfying the following four axioms:

1. If $e : A \longrightarrow B, e \in E$ and $e' : B \longrightarrow C, e' \in E$, then $ee' : A \longrightarrow C$, and $ee' \in E$.
Dually, if $m : A \longrightarrow B, m \in M$ and $m' : B \longrightarrow C, m' \in M$, then $mm' : A \longrightarrow C$, and $m'm \in M$.
2. If $e : A \longrightarrow B, e \in E$, then e is an epimorphism. Dually, if $m : A \longrightarrow B, m \in M$, then m is a monomorphism or a monic.
3. If $f : A \longrightarrow B$ is an isomorphism, then $f \in E \cap M$.
4. Every morphism $f : A \longrightarrow B$ in $\mathcal{A}$ has an E - M factorization which is unique up to an isomorphism, i.e., there exists an E - M factorization (e, m) of f , meaning that there exist $e \in E$ and $m \in M$ such that $f = em$. This results in the existence of the object $f(A)$ with $e : A \longrightarrow f(A)$ and $m : f(A) \longrightarrow B$.

Examples of such image factorization system are easy to construct. In the category of sets, for example, the set E of onto functions and the set M of one-to-one functions form

an image factorization system. This property of categories is an important ingredient of decision theory settings. Open categories cover all these requirements.

Definition 18 *A Category $\mathcal{A}$ is said to be **open** if it has an image factorization system $\{E, M\}$, pullbacks over the M -maps, and finite disjoint coproducts.*

Cockett uses the setting of open categories for discrete decision theory. The same setting is adopted in this dissertation for stochastic decision theory.

3.2 DISCRETE DECISION THEORY

This section is a review of the relevant work of Cockett pertaining to discrete decision theory [16,17,18] upon which this dissertation is based. The importance of this work stems from the fact that decision trees are formally viewed as terms of decision theories and, as a result, they can be manipulated algebraically. This refutes the long popular criticism of decision trees as being a primitive and “knowledgeless” representation of some process flow of control.

A decision theory is an equational theory whose models are coalgebras, instead of algebras. This means that its functions are cooperations, instead of operations. Cooperations can be used, in a particular context, to simulate decision processes. Suppose we are given two maps

$$tr : A \longrightarrow A, \text{ and } at : A \longrightarrow |n|,$$

where $|n|$ denotes a set of cardinality n . A map $\langle tr, at \rangle$ can be induced such that

$$\langle tr, at \rangle : A \longrightarrow A \times |n|.$$

As the object $A \times |n|$ is isomorphic to the n -fold disjoint union of A , there exists a

one-to-one map c_n such that

$$c_n : A \times |n| \longrightarrow A \oplus \cdots \oplus A,$$

making $\langle tr, at \rangle c_n$ a cooperation. The map at is referred to as an attribute map. The map tr is referred to by Cockett as a “twisting” endomorphism, and when it is the identity in A (later referred to as the idempotence property), the induced cooperation is called a *decision*. Its effect is similar to the one of a decision in a classification problem; it distributes a set of objects into a number of disjoint subsets. Cockett, in [18], derived the more general and important result that every cooperation can be uniquely factored into an induced decision followed by an endomorphism.

Since cooperations are the dual of operations, the basic methods of manipulation are analogous. For the sake of clarity, however, Cockett uses the notation $\langle x_1, \dots, x_n \rangle f$ for operations and the notation $f \langle x_1, \dots, x_n \rangle$ for cooperations. The advantage of formulating decision trees as terms of a decision theory is that their manipulations are induced by identities which govern the cooperations of the theory.

Definition 19 *A decision theory, $D = (\mathbf{Q}, \mathbf{E})$ consists of a set $\mathbf{Q}$ of cooperations and a set $\mathbf{E}$ of identities. Each $q_i \in \mathbf{Q}$ has an associated codomain, $\text{codom}(q_i) = \{a_1, a_2, \dots, a_{m_i}\}$. The identities of $\mathbf{E}$ are of the form $t_1 = t_2$, where t_1 and t_2 are terms of D .*

Terms are recursively defined as follows:

- (i) *There is a stock of variables $x_1, x_2, \dots$ each of which is a term.*
- (ii) *If $t_1, t_2, \dots, t_m$ are terms and q is a cooperation of arity m , then $q \langle t_1, t_2, \dots, t_m \rangle$ is a term.*

Every decision theory must satisfy the following identities for every $q, q_1, q_2 \in \mathbf{Q}$:

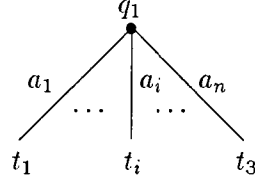
$$\begin{aligned} \text{[D.1]} \quad q\langle x' \cdots x \rangle &= x \\ &\text{(Idempotence)} \end{aligned}$$

$$\begin{aligned} \text{[D.2]} \quad q_1\langle x_1' \cdots x_{r-1}' q_2\langle y_1' \cdots y_n' \rangle x_{r+1}' \cdots x_m' \rangle \\ &= \\ q_2\langle q_1\langle x_1' \cdots x_{r-1}' y_1' x_{r+1}' \cdots x_m' \rangle \cdots q_1\langle x_1' \cdots x_{r-1}' y_n' x_{r+1}' \cdots x_m' \rangle \rangle \\ &\text{(Distribution)} \end{aligned}$$

$$\begin{aligned} \text{[D.3]} \quad q\langle x_1' \cdots x_{r-1}' q\langle y_1' \cdots y_n' \rangle x_{r+1}' \cdots x_m' \rangle \\ &= \\ q\langle x_1' \cdots x_{r-1}' y_r' x_{r+1}' \cdots x_m' \rangle \\ &\text{(Repetition)} \end{aligned}$$

A model of a decision theory $\mathbf{D} = (\mathbf{Q}, \mathbf{E})$ is a coalgebra (dual notion to algebra), called a *decision coalgebra*, in a category $\mathcal{C}$. Thus, a decision coalgebra is a base set A together with an interpretation of the cooperations of $\mathbf{Q}$ as decisions on A . Decision coalgebras, in turn, also form a category whose morphisms are morphisms of coalgebras which commute with each cooperation in the obvious manner. The category $\mathcal{C}$ must, however, be *open* (see [18] for more details). When, in addition, it *has a final object*, each decision interpretation corresponds to an attribute map. In this way, this setting tightly simulates decision processes. Each model of the theory is further required to satisfy the identities [D.1]–[D.3] and those in $\mathbf{E}$, when they exist, as well. In fact, a decision coalgebra must always satisfies the identities [D.1]–[D.3] and, in addition, the identities in $\mathbf{E}$ which characterize a particular decision theory. The identities [D.1]–[D.3] make decisions theories simulate our intuition about decision trees. Thus, given a cooperation

$q \in \mathbf{Q}$, with arity $n \geq 2$ and codomain $\text{codom}(q) = \{a_1, \dots, a_i, \dots, a_n\}$, the quantity $q\langle t_1, \dots, t_i, \dots, t_n \rangle$ denotes the following decision tree:



An important identity, transposition, is due to Chen and Ras [12] and is implied by the identities [D.1]–[D.3].

$$\begin{aligned}
 \text{[D.4]} \quad q_1 \langle q_2 \langle x_{11}, \dots, x_{1n} \rangle, \dots, q_2 \langle x_{m1}, \dots, x_{mn} \rangle \rangle \\
 = \\
 q_2 \langle q_1 \langle x_{11}, \dots, x_{m1} \rangle, \dots, q_1 \langle x_{1n}, \dots, x_{mn} \rangle \rangle \\
 \text{(Transposition)}
 \end{aligned}$$

This provides alternative definitions of decision theories as seen from the following lemma.

Lemma 1

- (i) The identities [D.2] and [D.3] imply [D.4].
- (ii) The identities [D.1] and [D.4] imply [D.2].

Proof:

See Cockett [16] for a complete proof. □

Cockett derived the key result to decision tree manipulations by showing that if every cooperation satisfies the identities [D.1]–[D.3], then so does every term of the decision theory. This result is a necessary generalization of the identities [D.1]–[D.3] for useful

manipulations to be carried out. Another important result, derived by Cockett, is the repeat reduction lemma which endorses the consensus of consistency in any decision making process, meaning that decisions previously made will be made in the same way as the process goes on.

Lemma 2 (The Repeat Reduction Lemma)

Let $D = (Q, E)$ be a decision theory with $q \in Q$ and w a term, then

$$\begin{aligned} q\langle x_1' \cdots x_{r-1}' w\langle y_1' \cdots y_{s-1}' q\langle z_1' \cdots z_n' \rangle \cdots y_m' \rangle \cdots x_n' \rangle \\ = \\ q\langle x_1' \cdots x_{r-1}' w\langle y_1' \cdots y_{s-1}' z_r' \cdots y_m' \rangle \cdots x_n' \rangle. \end{aligned}$$

Proof:

See Cockett [16] for a complete proof. □

This lemma and the identities [D.1]–[D.3] point to the existence, for every term of the decision theory, of “suitable” forms (in some sense), which are terms equivalent to it. The following definitions introduce interesting forms of decision theory terms and how one can produce them.

Definition 20

- (i) A term of a decision theory is said to be **repeat reduced** if the repeat reduction lemma cannot be applied to it.
- (ii) A term of a decision theory is said to be **idempotent reduced** if no cooperation can be eliminated by idempotence.
- (iii) A term of a decision theory is said to be **simply reduced** if it is both idempotent reduced and repeat reduced.

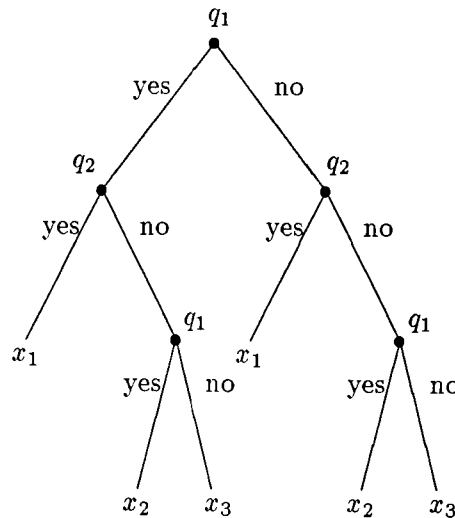
Lemma 3

- (i) Every term of a decision theory has a unique repeat reduced form.
- (ii) Every term of a decision theory has a unique idempotent reduced form.

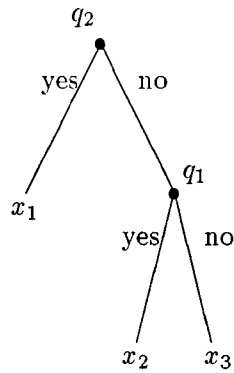
The idempotent reduction and the repeat reduction are not confluent operations for the simple reduction. A given term may have not have a unique simply reduced form. The following example illustrates this problem.

Example 7 (The First Simple Reduction Example)

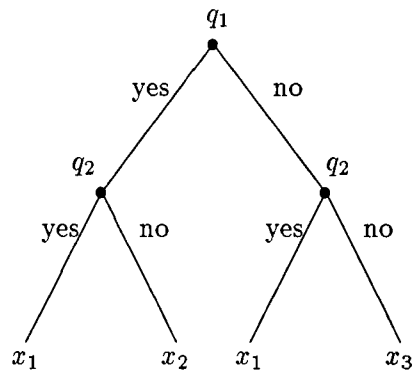
Consider the following decision tree:



Applying the idempotent reduction, the following simply reduced decision tree is obtained:



On the other hand, applying the repeat reduction yields the following simply reduced decision tree:

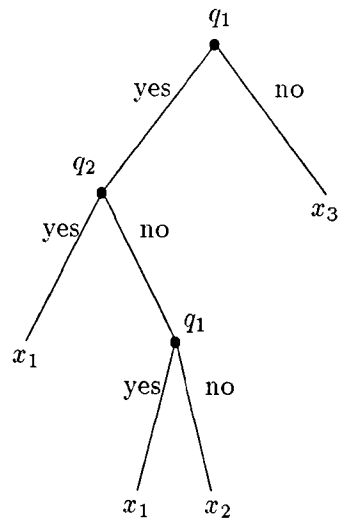


□

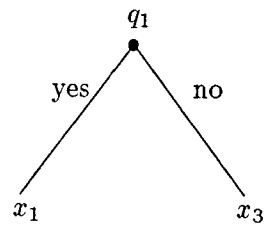
Simple reduction must be done in this order: repeat reduction followed by idempotent reduction. The reason for this is that if idempotent reduction is applied first, the repeat reduction which follows may produce a non-idempotent reduced term. The following example demonstrates this problem.

Example 8 (The Second Simple Reduction Example)

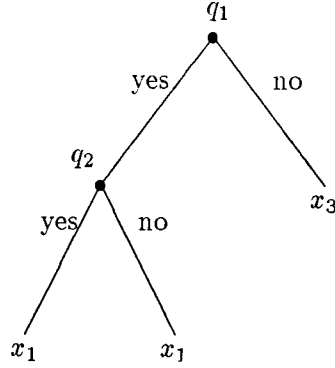
Consider the following decision tree:



Applying the simple reduction (repeat reduction followed by idempotent reduction), the following simply reduced decision tree is obtained:



Doing the reverse operation (idempotence reduction followed by repeat reduction), the following decision tree is obtained:



Obviously, this tree is not simply reduced; it further needs an idempotent reduction to become simply reduced.

□

One way to avoid the problems described in the previous examples is to define the simple reduction in this order: repeat reduction followed by idempotence reduction. The following uniqueness result then holds:

Lemma 4 *Every term of a decision theory has a unique simply reduced form.*

At this point, the important question whether the identities [D.1]–[D.3], introduced for the theory, are *complete* and *sound*. Completeness means that these identities can, alone, derive any other identity which holds in all the coalgebras of a given decision theory even when additional identities are present. Soundness means that there exists a non-trivial coalgebra which satisfies the identities [D.1]–[D.3]. In [18], Cockett proves these results in the general case and argues that decision trees can be viewed as coalgebraic entities whose manipulation is completely determined by the identities [D.1]–[D.3].

The manipulation of decision theory terms also raises the important question of equivalence between terms. There are several notion of equivalence for terms and one should distinguish between them.

Definition 21 Two terms, t_1 and t_2 , of a decision theory $\mathbf{D} = (\mathbf{Q}, \mathbf{E})$ are

- (i) **structurally identical**, denoted $t_1 \equiv t_2$, if and only if t_1 is the exact copy of t_2 .
- (ii) **decision equivalent**, denoted $t_1 =_d t_2$, if and only if they are equivalent with respect to the identities [D.1]–[D.3].
- (iii) **transposition equivalent**, denoted $t_1 =_t t_2$, if and only if they are equivalent with respect to the transposition identity alone.

To check that two terms are structurally identical, it suffices to compare the roots of the two terms and then proceed recursively to the subtrees at the leaves of the root decision. Checking decision equivalence, however, is a little harder. Two terms are decision equivalent if by using the identities [D.1]–[D.3] alone, one must be able to manipulate each term into the other. This does not provide us, though, with a method to check decision equivalence. The following technique does it, but at the expense of augmenting the size of the term. Suppose we want to check whether two terms t_1 and t_2 are decision equivalent. We first form the term t_1/t_2 obtained by appending to each leaf of t_1 a copy of the term t_2 . Denoting by $idp(t)$ the idempotent reduced form of the term t and by $rpt(t)$ its repeat reduced form, the following theorem claims that a necessary and sufficient condition for two terms t_1 and t_2 to be decision equivalent is that the simply reduced form of the term t_1/t_2 be structurally identical to the simply reduced form of the term t_1 .

Theorem 1 Two terms, t_1 and t_2 , of a decision theory $\mathbf{D} = (\mathbf{Q}, \mathbf{E})$ are [D.1]–[D.3] equivalent if $idp(rpt(t_1/t_2)) \equiv idp(rpt(t_1))$.

Proof:

See Cockett [16] for a complete proof.

□

Cockett also provided another alternative in deciding upon term equivalence. This alternative relies on path comparison. To this end, some definitions are necessary.

Given a decision q with the codomain $\text{codom}(q) = \{a_1, \dots, a_i, \dots, a_m\}$, a *branch instruction*, $q \langle ? \rangle a_i$ is an indication that a_i has been selected for decision q . A path in a decision tree starting from the root is completely determined by the list of branch instructions it follows until a leaf is reached. Such a path is called a *root path* and can be represented by lists of the form $[q_1 \langle ? \rangle a_1, \dots, q_n \langle ? \rangle a_n]$.

Definition 22

- (i) A **spine** of a term t is a pair $p(P, x)$ where $P = [q_1 \langle ? \rangle a_1, \dots, q_n \langle ? \rangle a_n]$ is the root path in t which leads to the variable x at some leaf of t . The set of all the spines of t is denoted by $Sp(t)$.
- (ii) A root path P is said to **agree** with a conjunct E of branch instructions if whenever $q \langle ? \rangle x$ or $q \langle ? \rangle y$ are in either the root path or the conjunct, then $x = y$.
- (iii) Given a root path P , the conjunct formed from its branch instruction is denoted by $\&P$.
- (iv) For a term t and a conjunct E , as above, the **retrieval** of E in t is defined as

$$\text{Ret}(E, t) = \{x \mid p(P, x) \in Sp(t) \text{ and } P \text{ agrees with } E\}.$$

The following theorem introduces another alternative test for decision equivalence between terms in decision theory.

Theorem 2 Given two terms, t_1 and t_2 of a decision theory, $t_1 =_d t_2$ if and only if for every spine $p(P, x)$ in t_1 , $\text{Ret}(\&P, t_2) = \{x\}$.

Proof:

See Cockett [16] for a complete proof. □

The problem of optimizing decision trees has been traditionally approached with respect to criteria measuring costs for evaluation, storage, and/or testing of these decision trees. As the information required by these methods is frequently not available, Cockett [17] proposed that the optimization be instead done with respect to a class of cost criteria, called *reasonable cost criterion*, which agrees with a *reasonable preorder*. The problem of optimizing a decision tree with respect to a reasonable criterion is that of finding, among the class of terms which are semantically equivalent to the original tree, a *minimal* element in the reasonable preorder (see [17,32,33] for more details). This minimal element is said to be irreducible. In [20], Cockett argued that for every irreducible decision tree, there exists a cost criterion with respect to which this decision tree is optimal. Conversely, given a cost criterion, the solution to the optimization problem is an irreducible decision tree. The optimization problem can then be solved by developing techniques to generate, for a given decision tree, all of its irreducible forms. For a given application, the selection of the best tree or set of trees will then be narrowed down to the set of irreducible forms. A characterization of irreducible forms is given in the following definition:

Definition 23 *A term t is said to be irreducible if all of its transposition equivalent terms are simply reduced.*

In [32,33], the authors described an algorithm for generating all the irreducible forms of a given term. This algorithm, for the sake of efficiency, must avoid the generation of transposes of irreducible forms. This constitutes a big saving, as the number of these transposes may be quite large. To solve this problem, the authors developed a technique

called *separation planning* which considerably reduces the search space and guarantees the generation of only representative irreducibles. Cockett [20] also developed a *reduction algorithm* for reducing a decision tree to an irreducible form. This algorithm has the property of generating irreducible terms which are always no-less efficient than the term they are derived from.

The notion of transposition equivalence between terms is very important to discrete decision theory. A method for determining transposition equivalence, due to Chen and Ras [12], states that

Theorem 3 *Two repeat reduced terms have bag-equivalent spine sets if and only if they are transposition equivalent.*

Proof:

See Cockett [16] for a complete proof.

□

The notion of bag-equivalence of two spine sets simply means that the conjunct of every root path in one spine set agrees with a conjunct of some root path in the other spine set and vice versa. In other words, all the root paths agree except for the order in which the branch instructions are encountered in the root paths.

The following definition gives an important characterization of the decisions in a term that allows these decisions to be transposed to the root of the term. This characterization, which Cockett names “semi-essentiality”, is particularly important, as will be shown in this dissertation, in stochastic decision theory.

Definition 24 *A decision q is said to be semi-essential in a term t if it occurs in the root path of every spine.*

Lemma 5 *A decision q is semi-essential in a term t if and only if there exists a term t' , transposition equivalent to t , which has the decision q at its root.*

Proof:

See Cockett [16] for a complete proof. □

3.3 MEASURE AND PROBABILITY THEORY

We adopt the intuitive concept of a *set* as a collection of objects, called *elements* or *members* of the set. The void set is denoted by $\emptyset$ and is a subset of every set. If S is a set, then the *power set* of the set S , $\mathcal{P}(S)$, is defined as the set of all subsets of S . The set theoretic operations of union ($\cup$), intersection ($\cap$), and complementation ($-$) have their usual meaning.

3.3.1 Definitions in Measure Theory

A **ring** (or **Boolean ring**) of sets is a non-empty class R of subsets of a set X such that if $E \in R$ and $F \in R$, then $E \cup F \in R$ and $E - F \in R$. In other words, a ring is a non-empty class of subsets of a set X which is closed under set union and set difference. While a ring is always closed under the intersection operation, a class of sets closed under intersection and difference is not necessarily a ring.

An **algebra** (or **Boolean algebra**) of sets is a non empty class R of subsets of a set X such that

(i) if $E \in R$ and $F \in R$, then $E \cup F \in R$, and

(ii) if $E \in R$, then $E' = X - E \in R$.

Every algebra is also a ring. An algebra may be characterized as a ring which contains the whole set X .

A σ -ring is a non empty class S of subsets of a set X such that

(i) if $E \in S$ and $F \in S$, then $E - F \in S$, and

(ii) if $E_i \in S, i = 1, 2, \dots$, then $\cup_{i=1}^{\infty} E_i \in S$.

In other words, a σ -ring is a ring closed under countable union. A σ -ring is also closed under countable intersection.

A σ -algebra of sets is a non empty class S of subsets of a set X such that

(i) if $E_i \in S, i = 1, 2, \dots$, then $\cup_{i=1}^{\infty} E_i \in S$, and

(ii) if $E \in S$, then $E' = X - E \in S$.

Similarly, a σ -algebra may be characterized as a σ -ring which contains the whole set X . If S is a σ -algebra in the set X , then X is said to be a **measurable space** and the members of S are called **measurable sets**.

A **positive measure** is a function μ , defined on a σ -algebra S , whose range is in $[0, \infty]$ and which is *countably additive*. This means that for a countable collection $\{A_i\}$ of disjoint members of S ,

$$\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i).$$

A **measure space** is a measurable space which has a positive measure defined on the σ -algebra of its measurable sets.

3.3.2 Probability Theory

Traditionally, probability theory deals with Boolean algebras of sets. An event is viewed as a set, and its opposite event is its complementary set; mutually exclusive events

are disjoint sets, and an event consisting of the simultaneous occurrence of two other events is the intersection set. This interpretation is adequate for translating physical terminologies into set theoretic terminologies for a class of simple problems. For other more useful situations, however, it is necessary to make an additional assumption to deal with countability: The system of events must be closed under the countable union, meaning that we must deal with a σ -algebra instead.

A **probability space** is a measure space (S, Ω, p) , where S is the set of **elementary events**, Ω is a σ -algebra of subsets of S , called the set of *possible* or *measurable events*, and p is a measure on S , mapping Ω into the unit interval $[0, 1]$ with $p(S) = 1$. The measure p is also called a **probability measure**.

Given some event $B \in \Omega$, with non zero probability, the **conditional probability** of an event A , given that event B has occurred, is given by

$$p(A | B) = \frac{p(A, B)}{p(B)}.$$

The probability of an event $A \in \Omega$, a σ -algebra in the set S , can be expressed in terms of the conditional probabilities of any partition of the set S into a number of mutually exclusive events. This is the **total probability theorem**. Let $B_n, n = 1, 2, \dots, N$, be the N mutually exclusive events whose union is the set S ; we have that

$$p(A) = \sum_{i=1}^N p(A | B_i) p(B_i).$$

Applying the definition of conditional probability to the event A and to one of the events defined above, say B_n , we can write that

$$p(A | B_n) = \frac{p(A, B_n)}{p(B_n)}, \text{ or } p(B_n | A) = \frac{p(B_n, A)}{p(A)}.$$

Combining these two equations leads to

$$p(B_n | A) = \frac{p(A | B_n) p(B_n)}{p(A)}.$$

This is one of the forms of **Bayes' theorem**. Another form is obtained by writing the total probability theorem for the event A . This leads to

$$p(B_n | A) = \frac{p(A | B_n)p(B_n)}{\sum_{i=1}^N p(A | B_i)p(B_i)}.$$

Independence between events is defined as follows:

Definition 25 *Two events A and $B \in \Omega$ are statistically independent [56] if and only if*

$$p(A, B) = p(A)p(B).$$

In terms of conditional probabilities, this condition is equivalent to say that

$$p(A|B) = p(A),$$

whenever $p(A|B)$ is defined ($p(B) \neq 0$).

The concept of independence can be extended to more than two events. For instance, the independence conditions for three events A , B , and $C \in \Omega$ are

$$\left\{ \begin{array}{lcl} p(A, B) & = & p(A)p(B) \\ p(A, C) & = & p(A)p(C) \\ p(B, C) & = & p(B)p(C) \\ p(A, B, C) & = & p(A)p(B)p(C). \end{array} \right.$$

Note that all the four conditions have to be simultaneously satisfied and that *pairwise independence is not sufficient* for the entire set of events to be *mutually independent*. Equivalently, the three events A , B , and C are *mutually statistically independent* if and

only if

$$\left\{ \begin{array}{lcl} p(A|B) & = & p(A) \\ p(A|C) & = & p(A) \\ p(B|C) & = & p(B) \\ p(A|B, C) & = & p(A) \\ p(C|A, B) & = & p(C) \\ p(B|C, A) & = & p(B), \end{array} \right.$$

whenever the conditional probabilities are defined.

In general, for n events $A_1, A_2, \dots, A_n \in \Omega$ to be *mutually statistically independent* [56], the following $2^n - (n + 1)$ equations must hold:

$$p(A_i, A_j, \dots, A_k) = p(A_i)p(A_j) \cdots p(A_k), \quad \forall i, j, \dots, k \leq n.$$

This is also equivalent to

$$p(A_i|A_j, \dots, A_k) = p(A_i), \quad \forall i, j, \dots, k \leq n.$$

In fact, these conditions are redundant and can be relaxed as it is shown in the following theorem.

Theorem 4 *Let $A_1, A_2, \dots, A_n \in \Omega$. The following statements are equivalent:*

- (i) *The n events are mutually statistically independent.*
- (ii) *Any ordering $l_1, l_2, \dots, l_n$ of these events satisfies the following $n - 1$ equations:*

$$\begin{aligned} p(A_{l_1}, A_{l_2}) &= p(A_{l_1})p(A_{l_2}) \\ p(A_{l_1}, A_{l_2}, A_{l_3}) &= p(A_{l_1})p(A_{l_2})p(A_{l_3}) \\ &\vdots \\ p(A_{l_1}, A_{l_2}, A_{l_3}, \dots, A_{l_n}) &= p(A_{l_1})p(A_{l_2})p(A_{l_3}) \cdots p(A_{l_n}), \end{aligned}$$

or, equivalently,

$$\begin{aligned} p(A_{l_2}|A_{l_1}) &= p(A_{l_2}) \\ p(A_{l_3}|A_{l_2}, A_{l_1}) &= p(A_{l_3}) \\ &\vdots \\ p(A_{l_n}|A_{l_{n-1}}, \dots, A_{l_3}, A_{l_2}, A_{l_1}) &= p(A_{l_n}). \end{aligned}$$

(iii) *There exists an ordering $l_1, l_2, \dots, l_n$ of these events which satisfies the above equations.*

Proof:

The proof is in Appendix A.1 on Page 155. □

The concept of *entropy*, as defined by Shannon [64], is a measure, in some sense, of how much *information* is “produced” by a given process. The term “information” is used here in a special sense that must not be confused with its ordinary usage. It is a measure of how uncertain one can be, in a random experiment, about the outcome, or equivalently, it is a measure of the information obtained when one performs the experiment.

Definition 26 *Suppose that, for a given random experiment with a probability space (S, Ω, p) , a partition $\mathcal{U}$ of events of the set of elementary events S is given by $\mathcal{U} = \{A_1, A_2, \dots, A_n\}$, with $A_i \in \Omega$ and $p(A_i) = p_i$. The **entropy** of this partition $\mathcal{U}$ is defined as the quantity*

$$H(\mathcal{U}) = - \sum_{i=1}^n p_i \log p_i.$$

As an example, the entropy of a two-event partition $\mathcal{U}$ with probabilities p and $q = 1 - p$ is $H(\mathcal{U}) = -p \log p - (1 - p) \log(1 - p)$. As a function of p , the entropy $H(\mathcal{U})$ reaches a maximum at $p = 1/2$ (and thus $q = 1/2$). This is expected as this particular distribution represents the case in which one is equally uncertain as to which outcome is more likely to

occur. The experiment, once performed, produces the maximum increment in information, over all p . On the other hand, if $p = 1$ and thus $q = 0$, meaning that the first event is certain, no information is produced by the experiment, and the entropy must reach an absolute minimum, which it does as $H(\mathcal{U}) = -1 \times \log(1) - 0 \times \log(0) = 0$. These results can be trivially extended to multiple events.

If, for a given random experiment, the probability distribution on a given partition $\mathcal{U}$ is totally or partially unknown and if an estimate of the missing information is desired, one choice is a probability distribution on this partition which assumes minimum a priori information for maximum entropy. This concept is embodied in the *maximum entropy principle* which states that if a probability distribution on a given partition of events of a random experiment is unknown, the probability distribution which maximizes the entropy of the random experiment subject to any known constraints is the optimal choice.

Entropy can also measure uncertainty in conditioned events. The following is a definition of conditional entropy.

Definition 27 *Given the conditions of Definition 26, and given an event $M \in \Omega$ with $p(M) \neq 0$, the conditional entropy of the partition $\mathcal{U}$ upon the event M is defined by*

$$H(\mathcal{U} | M) = - \sum_{i=1}^n p(A_i | M) \log p(A_i | M).$$

The conditional entropy $H(\mathcal{U} | M)$ measures the uncertainty about the partition $\mathcal{U}$ in the subsequence of trials in which M occurs.

Let $\mathcal{M}$ be another partition of events such that $\mathcal{M} = \{M_1, M_2, \dots, M_m\}$. The conditional entropy of the partition $\mathcal{U}$ upon every event M_j is

$$H(\mathcal{U} | M_j) = - \sum_{i=1}^n p(A_i | M_j) \log p(A_i | M_j).$$

The conditional entropy of the partition $\mathcal{U}$ upon the partition $\mathcal{M}$ is the following $p(M_j)$ -weighted average of $H(\mathcal{U} \mid M_j)$:

$$H(\mathcal{U} \mid \mathcal{M}) = - \sum_{j=1}^m p(B_j) H(\mathcal{U} \mid M_j).$$

The concept of conditional entropy turns out to be quite useful to stochastic decision theory. It not only provides a measure of information in decision processes, but also yields a simple completion algorithm when information is missing. Chapter 5 will make these points.

3.4 SUMMARY

This Chapter has introduced the basic concepts of category theory which are used all along Chapter 4 in setting a mathematical model for stochastic decision theory. A brief review of Cockett's discrete decision theory is also presented. This material has dictated the framework in which stochastic decision theory will be defined, the manipulation rules, and the completeness results of Chapter 4. This chapter has also reviewed basic notions of probability and measure theory and stressed the concept of statistical independence and entropy. These two concepts are generalized to stochastic decision theory in Chapter 5 and Chapter 6. This generalization is essential to the cluster decomposition method which we develop in Chapter 7.

CHAPTER 4

STOCHASTIC DECISION THEORY

This chapter parallels in many ways Cockett's development of discrete decision theory as presented in Chapter 3 in the sense that we seek a stochastic version of this theory. This stochastic nature, however, gives the theory a rather distinctive flavor and dictates a somewhat different setting. Section 4.1 gives the intuitive considerations which lead to the setting. The setting, for which much of the credit goes to Cockett [15], is then formally described in Section 4.2, while Section 4.3 formally defines stochastic decision theory and the manipulations of its terms. Section 4.4 introduces the problem of model satisfaction in stochastic decision theory and derives the soundness result. Section 4.5 discusses structural retrieval in stochastic terms and the equivalence between them. Section 4.6 introduces the general retrieval algorithm and derives the completeness result. Section 4.7 summarizes the results and gives some concluding remarks.

4.1 INTRODUCTION

The manipulations of idempotence, transposition, and repetition, introduced in discrete decision theory, can be used, when modified to incorporate the probabilistic information, to manipulate the terms of the stochastic decision theory. However, there are two major differences: First, there is no stochastic equivalent for the distribution rule (see Chapter 3), and second, in order to obtain a complete set of manipulations it is necessary to introduce an additional rule, called *indifference*. Furthermore, not all the rules can be

soundly applied in both directions: in particular, idempotence is a one-way rule, meaning that more examples or what we later call models will be valid after a one-way rule application. As a result, these manipulations must be regarded as *inference rules* on the stochastic terms.

The main problem is to show that the manipulations we propose are sound and complete. In order to achieve this, it is necessary to introduce a semantic setting in which the models of the theory shall lie. Clearly, the setting must somehow involve probability or, at least, measure. Furthermore, it must have disjoint coproducts and, in fact, be an open category as in [18]. The category of *description spaces* provides such a setting. This category is a slight generalization of the category of bounded measure spaces whose morphisms do not increase measure. In this category, however, products do not exist. Fortunately, coproducts do exist and there is an image factorization system of description space maps through measure preserving monics (the M -maps). These limitations, however inconvenient, still allow the theory to be defined, as the chosen category is open.

Once the setting is in place, it is necessary to describe how a *stochastic decision theory* may be interpreted within the setting. One is then in a position to show the completeness and soundness of the inference rules. To this end, we must investigate the structural retrieval properties of stochastic terms and their relationship to arbitrary retrievals.

4.2 THE SETTING

If we wish to develop a theory of *stochastic decisions*, it is necessary to know in which category the models of this theory will lie. As there is no uniform consensus concerning which is the correct category for interpreting stochastic events, we start by describing the category in which these models shall lie. Obviously, the choice of this category is

dictated by the theory we wish to develop; it therefore must capture the intuition behind the development.

Before introducing this category more formally, let us discuss the intuition we have of an object in this setting. An object consists of a set of *descriptions*, each of which has an associated *measure* [63]. These descriptions belong to a *Boolean algebra* [28,29] referred to as a *descriptive language*. A *description space* is a descriptive language together with a measure on its descriptions.

If two descriptions are logically disjoint, then the measure of the union of the descriptions must be the sum of the measures of the individual descriptions. The assignment of values to descriptions must then obey the usual rules of *additivity* in *measure spaces* [63].

In specifying a category, one must describe its objects or models and specify the maps between them. There are a number of choices available for specifying the maps, and each would result in a theory with a different character. Using the example of the counting measure as a guide, we note that set maps should not increase measure. This property is described in terms of the image of subsets: namely the number of elements in the image of a subset must never exceed the number in the subset itself. In order to mimic this set based image map in Boolean algebras it is necessary to consider *order preserving* maps which are *left-adjoints* (in the usual categorical sense, also see Definition 28 below) of a Boolean algebra map.

A further essential consideration concerns the intuition of when two description spaces are equivalent. If everything is in the same *proportion*, then all the statistical inferences we can make from the two spaces will be the same. Thus, we should really like for two such spaces to be equivalent.

We now formalize these ideas. The first concept we shall introduce is the adjoint map pair between Boolean algebras, which is necessary in order to be able to discuss the image

of a map. To this end, consider a Boolean algebra $\mathcal{B}$. There is a standard *partial ordering* ($\leq$) associated with the algebra given by:

$$\forall x_1, x_2 \in \mathcal{B}, \quad x_1 \leq x_2 \iff x_1 \cap x_2 = x_1.$$

A map $b : \mathcal{B}_1 \longrightarrow \mathcal{B}_2$, where $\mathcal{B}_1$ and $\mathcal{B}_2$ are Boolean algebras, is said to be *order-preserving* if

$$\forall x, y \in \mathcal{B}_1, \quad x \leq y \implies b(x) \leq b(y).$$

Definition 28 (Description Maps) *An order-preserving map, $b : \mathcal{B}_1 \longrightarrow \mathcal{B}_2$, between partially ordered Boolean algebras is a **description map** if there is a Boolean algebra morphism, $b^* : \mathcal{B}_2 \longrightarrow \mathcal{B}_1$ (i.e., a map which preserves the operations of meet, join, and negation) such that if $x_1 \in \mathcal{B}_1$ and $x_2 \in \mathcal{B}_2$ then $b(x_1) \leq x_2$ if and only if $x_1 \leq b^*(x_2)$. The map b^* is the **right-adjoint** of b and b is the **left-adjoint** of b^* .*

It is well-known that description maps uniquely determine and are uniquely determined by their right adjoints, the Boolean algebra morphisms. Boolean algebra morphisms of finite Boolean algebras always have corresponding left-adjoint which are description maps. In this case, the description map corresponds to the image map on the set of atoms, while the Boolean map corresponds to the inverse image map as was discussed for the counting measure above.

A *measure* on a Boolean algebra consists of a map, $m : \mathcal{B} \longrightarrow \mathcal{R}^+$, from the elements of the algebra to the non-negative real numbers such that, whenever b_1 and b_2 are disjoint elements of the Boolean algebra (that is $b_1 \cap b_2 = \perp$, with $\perp$ denoting the Boolean algebra empty set) then $m(b_1 \cup b_2) = m(b_1) + m(b_2)$.

Definition 29 (Description Space Maps)

*The **category of description spaces**, $\mathcal{C}$, has as objects the pairs $(\mathcal{B}, m)$, where $\mathcal{B}$ is*

a Boolean algebra and m is a positive measure on $\mathcal{B}$. A **description space map**, $b : (\mathcal{B}_1, m_1) \longrightarrow (\mathcal{B}_2, m_2)$, is a description map $b : \mathcal{B}_1 \longrightarrow \mathcal{B}_2$ which does not increase measure, i.e., a description map which satisfies the following property:

$$m_1(x) \geq m_2(b(x)), \forall x \in \mathcal{B}_1.$$

A description space is, despite the rather abstract definition, often a rather mundane object. The example below illustrates this.

Example 9 (Example of a Description Space)

Suppose that a poll is taken on a population of a hundred people. The poll consists of the following three yes-or-no questions or, to use this chapter terminology, decisions:

q_1 : Do you agree with the president's internal policy?

q_2 : Do you agree with the president's foreign policy?

q_3 : Do you plan to vote for the president in the next elections?

The following table gives the result of this poll. Clearly, the poll classifies the poll takers into eight groups. Each compartment in the table represents one of these groups.

		q_3	
		yes	no
q_1	yes	6	9
	no	11	14
	yes	18	17
	no	13	12

The description space for this example is $X = (\mathcal{B}, m)$, where the Boolean algebra $\mathcal{B}$ is the power set of the cartesian product of the possible answers (or *codomains*, written $\text{codom}(q_i)$ below) of the three decisions of the poll,

$$\mathcal{B} = \mathcal{P}(\text{codom}(q_1) \times \text{codom}(q_2) \times \text{codom}(q_3)),$$

where $\text{codom}(q_1) \times \text{codom}(q_2) \times \text{codom}(q_3) = \{\text{yes}, \text{no}\} \times \{\text{yes}, \text{no}\} \times \{\text{yes}, \text{no}\}$. Every element of the cartesian product

$$\underline{c} = (c_1, c_2, c_3) \in \text{codom}(q_1) \times \text{codom}(q_2) \times \text{codom}(q_3)$$

corresponds to a compartment of the table. For instance, the following element belongs to $\mathcal{B}$ and is formed by the two elements of the cartesian product $\underline{c}_1$ and $\underline{c}_2$ corresponding to the compartments with the numbers 11 and 6, respectively:

$$B = \{(\text{yes}, \text{no}, \text{yes}), (\text{yes}, \text{yes}, \text{yes})\}.$$

If the number associated with each element $\underline{c}$ of the cartesian product of the codomains is $\text{num}(\underline{c})$, this example gives the following measure on $B \in \mathcal{B}$:

$$m(B) = \sum_{\underline{c}_i \in B} \text{num}(\underline{c}_i).$$

□

The category does not have many limits: for example, products do not exist. There are, however, an albeit very limited set of pullbacks, which turn out to play a crucial role. The situation for colimits is slightly better: coproducts certainly exist. Clearly, the category of finite sets with the counting measure forms a rather special (full and faithful) sub-category of the category $\mathcal{C}$, within which things are exceptionally well-behaved.

In addition, there is an image factorization system of description space maps through *measure-preserving monics*. This is obtained by taking the Boolean algebra generated by

the descriptions, in the codomain, which are contained in an image of a description of the domain, and simply using the measure of the codomain. It is clear that the descriptions of the codomain which are contained in an image description must, in particular, be contained in the image of $\top$, the Boolean algebra whole set. Thus, the image of a description space map under this factorization is determined by the image of the $\top$ of the domain. Image factorization systems were defined in Chapter 3; their formal properties are described in [2,48]. The measure-preserving monics, which are the M-maps of the image factorization system, play a rather important role in the results we need for the development below.

Definition 30 *A measure-preserving monic, $g : (\mathcal{B}_1, m_1) \longrightarrow (\mathcal{B}_2, m_2)$ is a description space map, which preserves the measure, in the following sense, $\forall B \in \mathcal{B}_1$, $m_1(B) = m_2(g(B))$ and whose underlying description map is monic (faithful or injective) and hereditary (i.e., if $B \leq g(B_1)$, then $B = g(B'_1)$ for some $B'_1 \in \mathcal{B}_1$).*

Theorem 5 (Measure-Preserving Monic Factorization)

Any description space map $f : \mathcal{B}_1 \longrightarrow \mathcal{B}_2$ can be factored into $e(f)m(f)$, where $m(f)$ is a measure-preserving monic such that

- (i) *If $f = f_1 f_2$ and f_2 is a measure-preserving monic, then there is a unique comparison measure-preserving monic, k , such that $f_1 = e(f)k$ and $k f_2 = m(f)$,*
- (ii) *If $f_1 g_2 = g_1 f_2$, then there is a unique cross map, k , such that $e(f_1)k = g_1 e(f_2)$ and $km(f_2) = m(f_1)g_2$.*

The proof of this is straightforward. The result gives one form, in which an image factorization system can be characterized.

Example 10 (A Measure-Preserving Monic)

An example of measure-preserving monic can be given using example 9. For any sub-

algebra of the underlying Boolean algebra $\mathcal{B}$, we may define the same measure on the descriptions. However, it is not quite as simple as that: we must exhibit a right adjoint for the description inclusion.

Consider the sub-algebra $\mathcal{B}_0$ defined by $\mathcal{B}_0 = \mathcal{P}(\{\text{yes}\} \times \text{codom}(q_2) \times \text{codom}(q_3))$, in which the *branch instruction* of the decision q_1 is restricted to “yes”. A sub-description space $X_0 = (\mathcal{B}_0, m_0)$ may be defined with m_0 being the restriction of m to the sub-algebra $\mathcal{B}_0$. Let g be the inclusion map defined by

$$\begin{aligned} g: \mathcal{B}_0 &\longrightarrow \mathcal{B} \\ B &\longmapsto B, \end{aligned}$$

then the right adjoint, g^* , is given by

$$\begin{aligned} g^*: \mathcal{B} &\longrightarrow \mathcal{B}_0 \\ B &\longmapsto B \cap \top_{\mathcal{B}_0}. \end{aligned}$$

To show that the map g^* is right-adjoint to the map g , we need to show that

$$g(B_0) \leq B_1 \iff B_0 \leq g^*(B_1),$$

which is clearly the case. This shows that g is a description map. Clearly, as it uses the same measure, it is also a measure-preserving map. Thus, it is a description space map.

□

The pullbacks which exist are, not surprisingly, those which are over the measure-preserving maps.

Lemma 6 *The category $\mathcal{C}$ has pullbacks over measure-preserving monics, and the pullback of a measure-preserving monic is, in turn, a measure-preserving monic.*

Proof:

The proof is given in Appendix B.1 on Page 157. □

As mentioned above the category has finite coproducts. Actually, they are rather special coproducts as they are *disjoint*, that is the intersections of the embeddings are initial.

Lemma 7 *The category $\mathcal{C}$ has finite coproducts whose embeddings are measure-preserving monics and disjoint.*

Proof:

The proof is given in Appendix B.1 on Page 157. □

The *initial description space* has a single description which has measure zero. As there are no Boolean algebra maps from the final Boolean algebra to a non-final algebra, it follows that this is a *strict* initial object. It is also clear that the embeddings of a coproduct intersect in the initial description space. In fact, as coproducts are disjoint in Boolean algebras with description maps, they are disjoint in description spaces. This means that the category $\mathcal{C}$ is an *open category* in the sense defined by Cockett [18] and in this dissertation (Chapter 3, on Page 42).

4.3 STOCHASTIC DECISION THEORIES

Since the category $\mathcal{C}$ has finite coproducts, we may discuss *cooperations*, *decisions*, and *coalgebras* over this category. A *cooperation*, q , as introduced in Section 3.1 of Chapter 3, on an object (a description space) $X = (\mathcal{B}, m)$, is a map from X to its n -fold coproduct, with n being the arity of q . A *coalgebra* over this category is any description space together

with a collection of cooperations defined on it. A decision coalgebra is a coalgebra all of whose cooperations are idempotent or decisions.

Example 11 Continuing with example 9, the poll questions can be interpreted as decisions on the description space $X = (\mathcal{B}, m)$ as follows:

$$\begin{aligned} q_i : \mathcal{B} &\longrightarrow \mathcal{B} \oplus \mathcal{B} \\ B &\longmapsto (B_1, B_2), \end{aligned}$$

such that $B_1 = \{\underline{c} \in B \mid c_i = \text{yes}\}$ and $B_2 = \{\underline{c} \in B \mid c_i = \text{no}\}$.

As each copy of the coproduct space is just a sub-algebra of $\mathcal{B}$, the measure defined on the coproduct space is unchanged and $\forall B \in \mathcal{B}, \quad m(B) = m(B_1) + m(B_2)$.

For instance, the questions q_1 and q_2 of Example 9 are interpreted on the element $\{(\text{yes}, \text{yes}, \text{yes}), (\text{yes}, \text{no}, \text{yes})\}$ of the Boolean algebra $\mathcal{B}$ as follows:

$$q_1(\{(\text{yes}, \text{yes}, \text{yes}), (\text{yes}, \text{no}, \text{yes})\}) = (\{(\text{yes}, \text{yes}, \text{yes}), (\text{yes}, \text{no}, \text{yes})\}, \perp),$$

$$q_2(\{(\text{yes}, \text{yes}, \text{yes}), (\text{yes}, \text{no}, \text{yes})\}) = (\{(\text{yes}, \text{yes}, \text{yes})\}, \{(\text{yes}, \text{no}, \text{yes})\}).$$

□

Lemma 8 *Given an object S , a decision $q : S \longrightarrow S \oplus S \oplus \cdots \oplus S$, and a measure-preserving monic $r : R \longrightarrow S$, making R a subobject of S , there exists a unique induced decision on the subobject R , $q \mid r : R \longrightarrow R \oplus R \oplus \cdots \oplus R$, which makes the following diagram a pullback:*

$$\begin{array}{ccc} S & \xrightarrow{q} & S \oplus \cdots \oplus S \\ r \uparrow & & \uparrow r \oplus \cdots \oplus r \\ R & \xrightarrow{q \mid r} & R \oplus \cdots \oplus R \end{array}$$

Proof:

The proof is given in Appendix B.2 on Page 158.

□

The manipulations of decision coalgebras are the same as those described in Cockett [16,17] when the presence of measure is ignored. The theory becomes more interesting if the measure aspects are considered. It is the presence of measure that makes the stochastic theory look somewhat different.

Using categorical notation of Cockett [16,17] which we introduced in Chapter 3, we denote the terms as

$$q_1 \langle x_1 \mid q_2 \langle x_2 \mid x_3 \rangle \rangle,$$

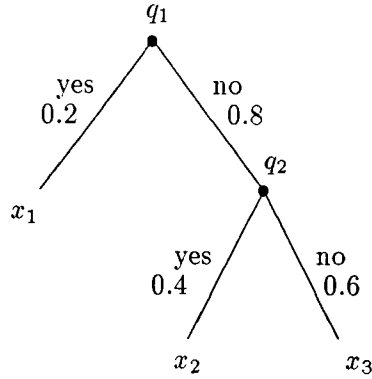
where “ $\langle x_2 \mid x_3 \rangle$ ” denotes a co-comparison map from the coproduct. However, as decisions often correspond to attribute maps, a more informative notation is to label the arguments by their attribute values:

$$q_1 : [yes \gg x_1, no \gg q_2 : [yes \gg x_2, no \gg x_3]].$$

This notation also allows us to label the branches of a decision by their probability of occurrence in the following fashion:

$$q_1 : [yes/0.2 \gg x_1, no/0.8 \gg q_2 : [yes/0.4 \gg x_2, no/0.6 \gg x_3]].$$

The set of labels or attributes associated with a decision is called the *codomain* of the decision. Stochastic terms can also be usefully graphically represented by a stochastic decision trees—the stochastic version of the familiar decision trees in which the branches are also labeled by the probability of occurrence. For instance, the previous stochastic term is represented by the following stochastic decision tree:



Definition 31 A stochastic decision theory, $SD = (Q, E)$ consists of a set Q of decisions and a set E of stochastic identities. Each q_i has an associated codomain, $\text{codom}(q_i) = \{a_1, a_2, \dots, a_{m_i}\}$. A decision's action is to select a member of the codomain based on tests to be performed upon known information.

Terms of stochastic decision theory are called **stochastic terms** and are recursively defined as follows:

- (i) There is a stock of variables $x_1, x_2, \dots$ each of which is a stochastic term.
- (ii) If $t_1, t_2, \dots, t_m$ are stochastic terms and q is a decision of arity m , then

$$q : [a_1/p_1 \gg t_1, \dots, a_m/p_m \gg t_m]$$

is a stochastic term with the p_i 's being probabilities satisfying $\sum_{i=1}^n p_i = 1$. This notation is equivalent to the statement that, for each $a_i \in \text{codom}(q)$, the probability of occurrence of a_i is p_i , and its occurrence results in the action t_i .

A **stochastic identity** is a stochastic term in only one variable. To distinguish between identities, we used the distinguished variable Ω , called the "full" variable.

In the next section, we shall describe how these terms are interpreted as decisions in the category of description spaces. Intuitively, however, the branching probabilities should

correspond to the conditional probabilities in the underlying description space model.

A stochastic term not only defines a decision, but also all the probabilities involved in that decision. It is clear that not every model will satisfy the constraints defined by the probability distribution. For this reason, stochastic terms act more like identities. Indeed, given a stochastic term, one can always form a stochastic identity by uniformly substituting the variables by the full variable Ω . In fact, we shall restrict our attention to stochastic decision theories in which $\mathbf{E}$ contains a single stochastic identity. Although this restriction still allows the derivation of the completeness result, it leaves unexplored the theories with many identities.

The manipulations of the terms of a discrete decision theory can be adapted to the stochastic case. This is done by incorporating the effect of the manipulations on the branching probabilities. Stochastic terms can then be manipulated in conjunction with their associated probability distributions. This constitutes an important merit of the stochastic manipulations, as they actually drive the changes in the branching probabilities.

The manipulations of stochastic terms are given by *inference rules*. An inference rule is of the form $t_1 \Rightarrow t_2$ and allows t_2 , with its variable substituted in some way, to be inferred whenever t_1 is valid. Until the notion of validity in this setting is formally introduced, these manipulations can, of course, only be regarded as formal devices.

The following are inference rules for stochastic terms:

For q , q_1 , and $q_2 \in \mathbf{Q}$, with arities m , m_1 , and m_2 , respectively, we have

$$[\text{SD.1}] \quad q : [a_1/p_1 \gg x, \dots, a_m/p_m \gg x] \Rightarrow x$$

(Idempotence)

$$[\text{SD.2}] \quad q : [a_1/p_1 \gg x_1, \dots, a_i/0 \gg x_i, \dots, a_m/p_m \gg x_m]$$

$$\Longleftrightarrow$$

$$q : [a_1/p_1 \gg x_1, \dots, a_i/0 \gg _, \dots, a_m/p_m \gg x_m]$$

where “ $_$ ” is any variable which has not occurred in the term

(Indifference)

$$[\text{SD.3}] \quad q : [a_1/p_1 \gg x_1, \dots, a_i/p_i \gg x_i, \dots, a_m/p_m \gg x_m]$$

$$\Longleftrightarrow$$

$$q : [a_1/p_1 \gg x_1, \dots, a_i/p_i \gg q : [a_1/0 \gg x_1, \dots, a_i/1 \gg x_i, \dots, a_m/0 \gg x_m], \dots, a_m/p_m \gg x_m]$$

(Repetition)

$$[\text{SD.4}] \quad q_1 : [a_{11}/p_1 \gg q_2 : [a_{21}/p_{11} \gg x_{11}, \dots, a_{2m_2}/p_{1m_2} \gg x_{1m_2}], \dots, a_{1m_1}/p_{m_1} \gg q_2 : [a_{21}/p_{m_11} \gg x_{m_11}, \dots, a_{2m_2}/p_{m_1m_2} \gg x_{m_1m_2}]]$$

$$\Longleftrightarrow$$

$$q_2 : [a_{21}/r_1 \gg q_1 : [a_{11}/r_{11} \gg x_{11}, \dots, a_{1m_1}/r_{1m_1} \gg x_{1m_1}], \dots, a_{2m_2}/r_{m_2} \gg q_1 : [a_{11}/r_{m_21} \gg x_{1m_2}, \dots, a_{1m_1}/r_{m_2m_1} \gg x_{m_1m_2}]]$$

(Transposition)

$$\text{with } \begin{cases} r_i &= \sum_{j=1}^{m_1} (p_j p_{ji}) \\ r_{ik} &= \frac{p_k p_{ki}}{\sum_{j=1}^{m_1} p_j p_{ji}} = \frac{p_k p_{ki}}{r_i} \end{cases}$$

where $i = 1, \dots, m_2$, and $j, k = 1, \dots, m_1$.

An example of how these inference rules may be used is provided by the proof of the following theorem.

Theorem 6 For any $q \in Q$,

$$[\text{SD.5}] \quad q : [a_1/0 \gg x_1, \dots, a_i/1 \gg x_i, \dots, a_m/0 \gg x_m] \implies x_i$$

(Certainty)

Proof:

The proof is in Appendix B.2 on Page 159. □

A *composite decision* is an encapsulation of a stochastic term and thus is equivalent to it. Any stochastic term can be encapsulated as a composite decision. The resulting composite decision will have as its codomain the set of variables of the stochastic term. The probability associated with each element of this codomain is then the probability of reaching the variable in the original term. When a stochastic term is so encapsulated, it becomes a black box the innards of which cannot be seen. This device is useful both in programming, as it is a subroutining facility, and as a theoretical tool.

The following examples illustrate how terms involving composite decisions may be formed from stochastic terms. Composite decisions shall be denoted by the letter w , rather than the letter q , which we shall reserve for primitive decisions.

Example 12 (Composite Decisions and Composite Terms)

If the composite decision w_1 is defined by

$$w_1 : [x_1/.08, x_2/.32, x_3/.18, x_4/.42]$$

$$\triangleq$$

$$q_1 : [a_1/.4 \gg q_2 : [b_1/.2 \gg x_1, b_2/.8 \gg x_2], a_2/.6 \gg q_3 : [c_1/.3 \gg x_3, c_2/.7 \gg x_4]],$$

and the composite decision w_2 by

$$w_2 : [x_1/.02, x_2/.44, x_3/.54]$$

$$\triangleq$$

$$q_1 : [a_1/.1 \gg q_2 : [b_1/.2 \gg x_1, b_2/.8 \gg x_2], a_2/.9 \gg q_3 : [c_1/.4 \gg x_2, c_2/.6 \gg x_3]],$$

one may form the composite term

$$w_1 : [x_1/.08 \gg w_2 : [x_1/.02 \gg y_1, x_2/.44 \gg y_2, x_3/.54 \gg y_3],$$

$$x_2/.32 \gg y_4, x_3/.18 \gg y_3, x_4/.42 \gg y_5],$$

in which the composite decisions w_1 and w_2 stand for the respective stochastic terms to which they are equivalent.

□

The following result says that composite decisions may be handled in exactly the same way as primitive decisions are handled. This is an essential ingredient of a sensible subroutining facility.

Theorem 7 *Composite decisions satisfy [SD.1] – [SD.4].*

Proof:

The proof is by induction on the height of these composite terms. The **height** of a composite term is defined as the number of decisions in the longest path in the term. The complete proof is given in Appendix B.2 on Page 160.

□

The following corollary uses composite decisions as a tool in deriving a useful manipulation, called *deep repetition*, which allows the elimination or introduction of a repeated decision at an arbitrary height above the first occurrence.

Corollary 1 (Deep Repetition Lemma)

$$\begin{aligned}
 \text{[SD.6]} \quad & q : [a_1/p_1 \gg t_1, \dots, a_i/p_i \gg w : [x_1/r_1 \gg X_1, \dots, x_j/r_j \gg X_j, \dots \\
 & \dots, x_n/r_n \gg X_n], \dots, a_m/p_m \gg t_m] \\
 & \iff \\
 & q : [a_1/p_1 \gg t_1, \dots, a_i/p_i \gg w : [x_1/r_1 \gg X_1, \dots, x_j/r_j \gg q : [a_1/0 \gg t_1, \dots \\
 & \dots, a_i/1 \gg X_j, \dots, a_m/0 \gg t_m], \dots, x_n/r_n \gg X_n], \dots, a_m/p_m \gg t_m] \\
 & \text{(Deep Repetition)}
 \end{aligned}$$

Proof:

The complete proof of this corollary is given in Appendix B.2 on Page 165. □

This allows some of the manipulations introduced for terms in a discrete decision theory to be generalized to stochastic terms. However, in introducing the generalizations, it is necessary to distinguish between the manipulations which give an equivalence of stochastic terms and those which give only a *one-way* inference.

For instance, applying the idempotence manipulation, which is only a *one-way* operation, repeatedly, until no idempotent sub-term is left, will result in an *idempotent-reduced* term. Every stochastic term has a *unique* idempotent-reduced inferred term. Notice, however, that the term obtained is not equivalent to the original.

The use of (deep) repetition, which is a *two-way* manipulation, allows the introduction of the analogue of the discrete notion of repeat reduction (see Cockett [16]). This reduces the tree size by removing previously occurring decisions. Applying this repeatedly to a stochastic term, until it is no longer possible, is of obvious utility and produces a *repeat-reduced* stochastic term. This process is called *repeat-reduction* and results in an equivalent stochastic term.

The *two-way* indifference manipulation is also important. If it is repeatedly applied in the direction which reduces the size of the term, thus introducing the "don't care variable" "-", an *indifference-reduced* term is obtained. A term which is both indifference-reduced and repeat-reduced is said to be a *reachable-reduced* term. The notion of reachability shall be important in the sequel and is intuitively connected to the fact that, in a stochastic term, only parts with non-zero probability can be visited. In the process of obtaining a reachable-reduced term, the order in which the manipulations are performed is unimportant, as these manipulations are indeed confluent and often reinforce each other. As shall be seen, the indifference manipulation can also be very useful in introducing non semi-essential decisions on the paths on which they are missing, thus allowing other manipulations, such as the reachable transposition, to be carried out.

The application of deep repetition can be considered to be equivalent to the one-way application of certainty. However, it is important to note that a *certainty reduction*, resulting in a *certainty-reduced* term, will not give an equivalent term. These rather subtle differences are in practice quite important, as it is often highly desirable to maintain *equivalence*.

As it will be seen in the next section, if the stochastic identity t_1 can be inferred from a stochastic identity t_2 , then the set of models which satisfy t_1 includes those which satisfy t_2 . However, If the inference rule is only one-way, then one of the main results of this chapter shows that the inclusion is proper. Thus, equivalence of terms is the assurance that one is handling the same class of models.

4.4 MODELS OF STOCHASTIC DECISION THEORIES

The models of a stochastic decision theory lie in the category of description spaces which we defined in section 4.3. This section will address the issues of model satisfaction and consistency.

A model M of a stochastic decision theory $\mathbf{SD}$ consists of an object X in the category $\mathcal{C}$ and for each $q_i \in \mathbf{Q}$, the set of decisions of the theory, an appropriate interpretation as a decision on X ,

$$M(q_i) : X \longrightarrow X \oplus X \oplus \cdots \oplus X,$$

such that each stochastic identity of $\mathbf{E}$ is *satisfied*. The notion of satisfaction is intuitively simple, namely that all the conditional probabilities must be correct. Stating this formally is more difficult. To simplify the notation, $M(q_i)$ will be simply written as q_i .

Definition 32 *Let M be an interpretation of the decisions of $\mathbf{Q}$ on an object X of the category $\mathcal{C}$, then X satisfies (or is consistent with) a stochastic term t if*

(i) t is a variable, or if

(ii) $t = q : [a_1/p_1 \gg t_1, \dots, a_n/p_n \gg t_n]$ and for each $i = 1, \dots, n$, the decision coalgebra

X_i induced by the map $\pi_i : X_i \longrightarrow X$ and given in the following pullback:

$$\begin{array}{ccc} X & \xrightarrow{q} & X \oplus \cdots \oplus X \\ \pi_i \uparrow & & \uparrow b_i \\ X_i & \xrightarrow{\pi_i} & X \end{array}$$

is such that

(a) X_i satisfies t_i , and

(b) $m_X(\top_X)p_i = m_{X_i}(\top_{X_i})$, where $\top_X$ and $\top_{X_i}$ are the “top” of the Boolean algebras $\mathcal{B}_X$ with the measure m_X and $\mathcal{B}_{X_i}$ with the measure m_{X_i} , respectively.

Notice that an interpretation into a description space with zero measure will satisfy every identity. Furthermore, expanding an interpretation using an expansion functor will not affect the validity of the interpretation.

Theorem 8 (Soundness of the Inference Rules)

Given a model M of a stochastic decision theory $\mathbf{SD}$ and an object X of the category $\mathcal{C}$, if X is consistent with a stochastic term t , then applying [SD.1] – [SD.4] results in a consistent stochastic term.

Proof:

The proof is given in Appendix B.3 on Page 167. □

Definition 33 *A stochastic term is said to be **repeat-consistent** if for any repeated decision, the probabilities are zero except on the appropriate argument or the path to the repeated decision has zero probability.*

Example 13 (Repeat-Consistent Stochastic Terms)

(1) The stochastic term

$$q_1 : [a_1/.3 \gg x_1, a_2/.4 \gg q_2 : [b_1/.2 \gg q_1 : [a_1/0 \gg x_2, a_2/1 \gg x_3, \\ a_3/0 \gg x_4], b_2/.8 \gg x_5], a_3/.3 \gg x_6]$$

is repeat-consistent.

(2) The stochastic term

$$q_1 : [a_1/.3 \gg x_1, a_2/.4 \gg q_2 : [b_1/0 \gg q_1 : [a_1/.6 \gg x_2, a_2/.1 \gg x_3, \\ a_3/.3 \gg x_4], b_2/1 \gg x_5], a_3/.3 \gg x_6]$$

is repeat-consistent.

(3) The stochastic term

$$q_1 : [a_1/.3 \gg x_1, a_2/.4 \gg q_2 : [b_1/.2 \gg q_1 : [a_1/.6 \gg x_2, a_2/.1 \gg x_3, \\ a_3/.3 \gg x_4], b_2/1 \gg x_5], a_3/.3 \gg x_6]$$

is not repeat-consistent.

□

Definition 34

(i) A description space $X = (\mathcal{B}, m)$ is said to be **full** if

$$\forall B \in \mathcal{B}, m_X(B) = 0 \implies B = \perp.$$

(ii) A description space $X = (\mathcal{B}, m)$ is said to be **non-trivial** if $m_X(\top_X) > 0$.

A set with the counting measure, is an example of a full and non-trivial description space.

Theorem 9 (Consistency of Repeat-Consistent Stochastic Terms)

A stochastic term t is repeat-consistent if and only if there exists a model over a non-trivial full description space which satisfies t .

Proof:

The proof is given in Appendix B.3 on Page 170.

□

Thus, the only stochastic terms which are satisfied by non-trivial models are the repeat-consistent terms. The remainder of the chapter is, thus, concerned with the properties of repeat consistent terms.

4.5 STRUCTURAL RETRIEVAL AND EQUIVALENCE

Given any stochastic term, one may write a path from root to leaf as a list of *branch instructions*. With each such path, one can associate the variable at the leaf and the probability of the path. Such an association is called a *stochastic spine* of the term and is denoted

$$\text{path}([q_1 \langle ? \rangle a_1, q_2 \langle ? \rangle a_2, \dots, q_n \langle ? \rangle a_n], x/p(x)),$$

where “ $\langle ? \rangle$ ” is the *branch operator* which associates a decision with a specified element of its codomain. The set of all stochastic spines with a *non-zero* probability, of a stochastic term t , is referred to as the *reachable stochastic spine set* of the term and is denoted $Sp(t)$. A *bag* is a list in which the order is unimportant. Two lists of branch instructions are *bag-equivalent* if one can be re-ordered into the other.

Definition 35

- (i) A decision is said to be **semi-essential** in a stochastic term t if it occurs in every path of t .
- (ii) A decision is said to be **reachable semi-essential** in a stochastic term t if it occurs in every non-zero probability path of t .

The notion of reachability is quite important in the sense that when a decision is missing in a zero probability path, we can always assume that it is present in view of the indifference rule [SD.2]. Clearly, every semi-essential decision of a stochastic term t is a reachable semi-essential decision; but the converse is not necessarily true. The following lemma demonstrates the utility of the concept.

Lemma 9 (Characterization of Reachable Semi-Essentials)

If a decision q is reachable semi-essential in a stochastic term t , then there is a stochastic term t' , rooted at q , such that $t \Longleftrightarrow t'$.

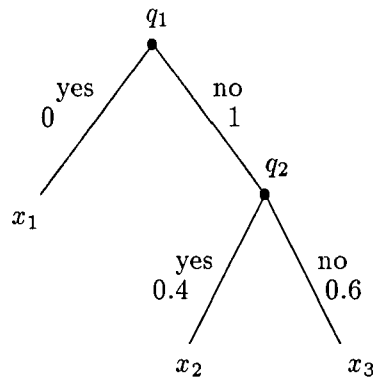
Proof:

The proof is given in Appendix B.4 on Page 172. □

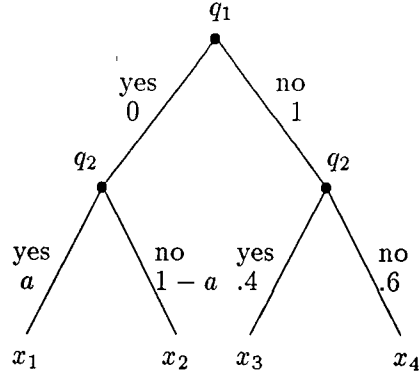
The previous lemma shows how the transposition rule [SD.4] may be applied in conjunction with the indifference rule [SD.2] to change the order of decisions in the term. When this technique is used, we obtain stochastic terms which are *reachable transposition equivalent*. Clearly, the effect of the usual transposition and the reachable transposition on the stochastic terms is very similar, especially if we only consider the reachable parts of the term.

Example 14 (Reachable Semi-Essential Decisions)

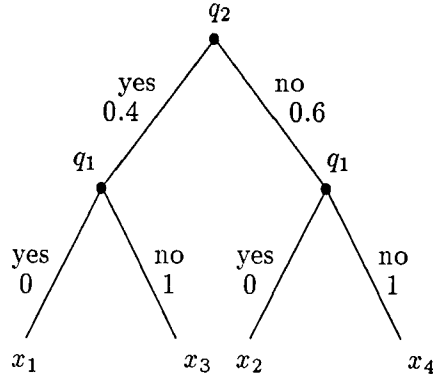
Consider the following stochastic term t :



Clearly, the decision q_2 is not semi-essential in t , but since $p(q_1 \text{ yes}) = 0$, q_2 is reachable semi-essential in t . By the indifference rule [SD.2], the stochastic term t is equivalent to the stochastic term t_1



where the decision q_2 is inserted at the impossible path of t . The decision q_2 is now semi-essential in the stochastic term t_1 , and the transposition rule [SD.4] can thus be used to transpose q_2 at the root of the stochastic term t_1 . The stochastic term t_2 , equivalent to t_1 is obtained and is



The stochastic terms t and t_2 are reachable transposition equivalent.

□

Definition 36 Two stochastic terms t_1 and t_2 are **stochastic spine set bag-equivalent** if there is an isomorphism of sets

$$\alpha : Sp(t_1) \longrightarrow Sp(t_2)$$

$$path(P_1, x_1/p(x_1)) \longmapsto path(P_2, x_2/p(x_2))$$

such that P_1 is bag-equivalent to P_2 , $x_1 = x_2$, and $p(x_1) = p(x_2)$.

Theorem 10 (Stochastic Version of Chen and Ras' Lemma)

Two repeat consistent repeat-reduced stochastic terms are stochastic spine set bag-equivalent if and only if they are reachable transposition equivalent.

Proof:

The complete proof is given in Appendix B.4 on Page 172.

□

A *conjunction* is a sequence of branch instructions and has the following form

$$S = q_1 \langle ? \rangle a_{i_1} \ \& \ q_2 \langle ? \rangle a_{i_2} \ \& \ \dots \ \& \ q_n \langle ? \rangle a_{i_n}.$$

Two conjunctions S_1 and S_2 are *logically equivalent* ($S_1 \iff S_2$) if using the usual identities of $\&$ (associativity, commutativity, idempotence) one can infer one conjunction from the other. The empty conjunction is logically always true and is denoted by $\top$.

A *root path* in a tree is a path starting at the root of the tree and working down towards the leaves. An occurrence of a subtree is determined by its root path. Given a path P , the conjunction of its branch instructions is denoted by $\&P$.

Definition 37 (Structural Retrieval)

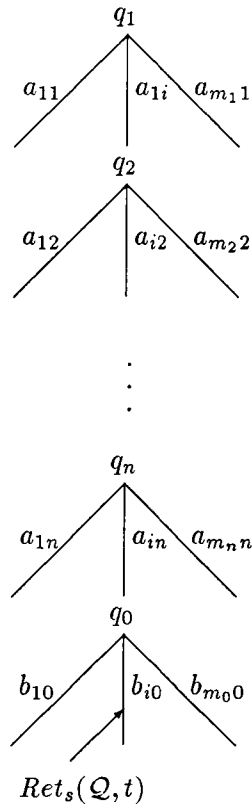
- (i) A **query**, $\mathcal{Q} = (q_0 \langle ? \rangle a_0 \mid C)$, is a pair where $q_0 \langle ? \rangle a_0$ is called the **event** and C is called the **condition**.
- (ii) A **structural retrieval tree** for $\mathcal{Q} = (q_0 \langle ? \rangle a_0 \mid C)$ has a reachable root path P such that $\&P$ is equivalent to the condition of the query and the subtree determined by P has the event decision as its root decision. The **structural retrieval**, denoted by $Ret_s(\mathcal{Q}, t)$ is the branch probability of the event in the subtree determined by P .

(iii) A query Q can be **structurally retrieved** in a repeat consistent tree t , if there is a tree t' , with $t \Rightarrow t'$, where t' is a structural retrieval tree for Q .

For instance, in order to structurally retrieve the query

$$Q = (q_0 <?> a_0 \mid q_1 <?> a_{i_1} \ \& \ q_2 <?> a_{i_2} \ \& \ \dots \ \& \ q_n <?> a_{i_n}),$$

in a given stochastic decision term t , the latter must be manipulated into a structural retrieval tree t' which is equivalent to t and has the following form:



The ordering in which the decisions of the conjunction appear in the root path of the term does not matter.

Of course, it will not always be the case that a structural retrieval can be performed. Also, as we shall see in the next section, there are retrievals which can be performed even

when a structural retrieval fails. However, the structural retrievals of a stochastic tree are important as, not only do they characterize the tree, but also they underlie the more complex retrievals introduced in the next section.

All retrievals correspond to conditional probability computations. Structural retrievals, in particular, do so in a very straightforward way as the root path gives the condition. This makes them an attractive vehicle for these computations.

For notational convenience, the "&" separator in conjunctions is dropped in favor of the more usual separator ",", for probabilistic events. Furthermore, as the retrieval of Q is the conditional probability of the event $q_0<?>b_{i_0}$ given C in a model,

$$Ret_s(Q, t) = p(q_0<?>b_{i_0} \mid C),$$

when the term in which the retrieval is being performed is clear from the context, we shall use the usual conditional probability notation for the retrieval.

Example 15 (Structural Retrieval)

Considering the poll of Example 9, suppose that the person conducting the poll acted improperly in two ways. Firstly, he asked the poll questions in an unusual order. Secondly, he assumed that if a person is not going to vote for the president in the next elections and yet agrees with the president foreign policy, this person cannot agree with the president's internal policy. The poll, as it was conducted, produced the stochastic decision tree, t , of Figure 4.1.

Suppose that, based on this poll, incomplete as it is, we would like to know what percentage of the population, who like the presidents internal policy but dislike his foreign policy, might still vote for him in the next elections. This can be expressed as the following query,

$$Q = (q_3<?>Yes \mid q_1<?>Yes, q_2<?>No).$$

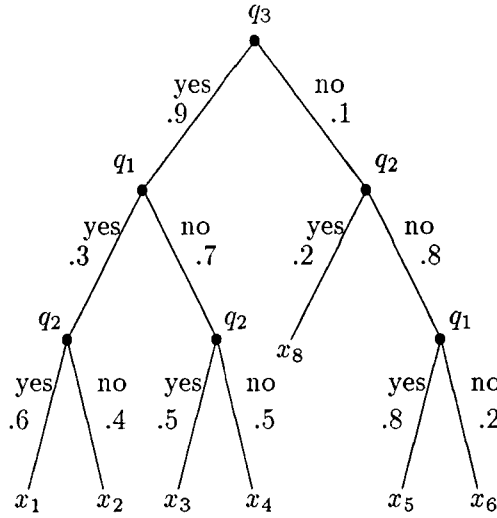


Figure 4.1: The Stochastic Decision Tree t of the Retrieval Example.

To answer this query, Q must be retrieved in t and the stochastic decision tree t must be manipulated into one in the required form. The resulting stochastic decision tree t' is depicted in Figure 4.2. From this tree, the retrieval of the query Q is clearly

$$p(q_1 \langle ? \rangle a_1 \mid q_2 \langle ? \rangle b_2, q_3 \langle ? \rangle c_1) = 0.89.$$

This means that 89% of the population, which agree with the president's internal policy yet disagree with his foreign policy, will still vote for the president in the next elections.

□

Structural retrievals are conceptually very clean as they may be driven by the structural form which must be achieved. Thus, the usual conditional probability computations can be entirely viewed in terms of some structural manipulations of stochastic terms. This gives a very practical method for doing some conditional probability calculations.

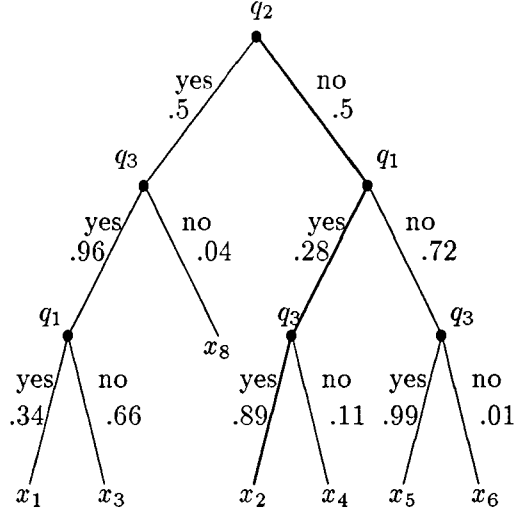


Figure 4.2: The Stochastic Decision Tree t' Under the Structural Retrieval Procedure.

The result, we now prove, although it is quite simple, has rather important ramifications. It is simply that in “computing” a retrieval by manipulating the tree into a structural retrieval tree, we may restrict our attention to the two-way manipulations.

Lemma 10 *A query Q can be structurally retrieved in a stochastic term t if and only if there exists a stochastic term t' such that $t \iff t'$ and t' is a structural retrieval tree.*

Proof:

The proof is given in Appendix B.4 on Page 173.

□

The fact that the manipulations can be chosen from the two-way inferences is useful as these manipulations preserve the reachable semi-essential properties of decisions of repeat consistent terms. This can be seen from an inspection of the rules. We use this fact

to obtain the following result which also holds for the more complex notion of retrieval introduced in the next section.

Lemma 11 *Given a query*

$$\mathcal{Q} = (q_0 <?> b_{0i_0} \mid q_1 <?> a_{1i_1}, q_2 <?> a_{2i_2}, \dots, q_n <?> a_{ni_n})$$

and a repeat consistent stochastic decision tree t such that

$$t \iff t_0 = q_1 : [a_{11}/p_{11} \gg t_1, \dots, a_{1i_1}/p_{1i_1} \gg t_{i_1}, \dots, a_{1n}/p_{1n} \gg t_n],$$

$\mathcal{Q}$ can be retrieved in t if and only if the query

$$\mathcal{Q}' = (q_0 <?> b_{0i_0} \mid q_2 <?> a_{2i_2}, \dots, q_n <?> a_{ni_n})$$

can be retrieved in the subtree t_{i_1} of t_0 and $Ret_s(\mathcal{Q}, t) = Ret_s(\mathcal{Q}', t_{i_1})$.

Proof:

The proof is given in Appendix B.4 on Page 174.

□

If a query has an empty condition then it is clear that the decision of the event must be reachable semi-essential in the term if it is to be structurally retrieved. This observation is crucial to the main result of this section which uses the structural retrievals in one tree to force a second tree into the same shape.

Theorem 11 (Characterization of Inference by Structural Retrieval)

For two repeat consistent stochastic identities t_1 and t_2 , $t_1 \implies t_2$ if and only if every query $\mathcal{Q}$ which can be structurally retrieved in t_2 can be structurally retrieved in t_1 and $Ret_s(\mathcal{Q}, t_1) = Ret_s(\mathcal{Q}, t_2)$.

Proof:

The proof is given in Appendix B.4 on Page 175.

□

This provides a method for determining the equivalence of two repeat consistent stochastic decision trees. They could be tested for equivalence by testing their structural retrievals, although that might be time consuming. The implied technique of rearranging one tree into the other is much more promising and may provide a tractable algorithm to determine stochastic term equivalence and whether one stochastic term implies another.

4.6 GENERAL RETRIEVAL AND COMPLETENESS

In this section, we wish to show that the retrieval properties of a stochastic term are equivalent to the computation of conditional probabilities on the non-trivial models. Thus, if all we know is that the identity holds, the only conditional probabilities we may compute, independently of further information on the model, are those which can be retrieved from the identity. Once this is established, it is easy to show that the identities are complete.

The retrieval of a query is a conditional probability computation. The structural retrievals only cover those conditional probability computations that can be performed by a structural manipulation of a stochastic term. It is therefore necessary to extend our notion of structural retrieval to the notion of general retrieval, even though we shall still use the structural retrievals, as they play an important role in the results.

Example 16 (The Failure of a Structural Retrieval)

Consider the retrieval of the query $(q_0 <?> yes \mid q_1 <?> no, q_2 <?> yes)$ in the following

stochastic term:

$$q_3 : [yes/0.2 \gg q_1 : [yes/0.5 \gg x_1, no/0.5 \gg q_2 : [yes/0.25 \gg q_0 : [yes/0.3 \gg x_2, \\ no/0.7 \gg x_3], no/0.75 \gg x_4]], no/0.8 \gg q_2 : [yes/0.4 \gg q_1 : [yes/0.3 \gg x_4, \\ no/0.7 \gg q_0 : [yes/0.6 \gg x_2, no/0.4 \gg x_3]], no/0.6 \gg x_1]].$$

The query cannot be retrieved by a structural retrieval, as neither q_1 nor q_2 are reachably semi-essential. However, since the query appears in every reachable subterm below the arguments of the first decision (q_3), the conditional probability can still be computed. The following development shows the derivation:

$$\begin{aligned} p(q_0 <?> yes \mid q_1 <?> no, q_2 <?> yes) &= \frac{p(q_0 <?> yes, q_1 <?> no, q_2 <?> yes)}{p(q_1 <?> no, q_2 <?> yes)} \\ &= \frac{p(q_0 <?> yes, q_1 <?> no, q_2 <?> yes, q_3 <?> yes) + \dots}{p(q_1 <?> no, q_2 <?> yes, q_3 <?> yes) + \dots} \\ &\quad \frac{\dots + p(q_0 <?> yes, q_1 <?> no, q_2 <?> yes, q_3 <?> no)}{\dots + p(q_1 <?> no, q_2 <?> yes, q_3 <?> no)} \\ &= \frac{0.3 \times 0.5 \times 0.25 \times 0.2 + 0.6 \times 0.7 \times 0.4 \times 0.8}{0.5 \times 0.25 \times 0.2 + 0.7 \times 0.4 \times 0.8} = 0.57. \end{aligned}$$

□

Definition 38 Let $C = q_1 <?> a_1 \ \& \ q_2 <?> a_2 \ \& \ \dots \ \& \ q_n <?> a_n$ be a conjunct and t a repeat-consistent stochastic term, then

1. The **retrieval** of the conjunction C in the stochastic decision term t is denoted

$Ret(C, t)$ and has its existence and value recursively defined as follows:

[Ret.1] If $C = \top$, then $Ret(C, t)$ exists and has value 1.

[Ret.2a] If $t = q_i : [\dots, a_i/0 \gg t', \dots]$, where q_i appears in C , then $Ret(C, t)$ exists and has value 0.

[Ret.2b] If $t = q_i : [\dots, a_i/p_i \gg t', \dots]$, where q_i appears in C and $p_i \neq 0$, then $Ret(C, t)$ exists and has value $Ret(C, t) = p_i Ret(C', t')$ if and only if the **residual retrieval** $Ret(C', t')$, with C' being C diminished by $q_i <?> a_i$, exists.

[Ret.3] If $t = q_0 : [a_{01}/p_{01} \gg t_1, \dots, a_{0m}/p_{0m} \gg t_m]$, where q_0 does not occur in C , then $Ret(C, t)$ exists and has value $\sum_{j \in J} p_{0j} Ret(C, t_j)$ if and only if for every $j \in J = \{ \text{indices for which } p_{0j} \neq 0 \}$, the residual retrieval $Ret(C, t_j)$ exists.

[Ret.4] If $t = X$, a variable, and $C \neq \top$, then $Ret(C, t)$ fails.

2. The **conditional retrieval** of the branch instruction $q <?> a$ given the conjunct C exists whenever $Ret(q <?> a \ \& \ C, t)$ and $Ret(C, t)$ exist and is **well-conditioned** if $Ret(C, t) \neq 0$, in which case it has a value

$$Ret(q <?> a \mid C, t) = \frac{Ret(q <?> a \ \& \ C, t)}{Ret(C, t)}.$$

If $Ret(C, t) = 0$, then the conditional retrieval is said to be **ill-conditioned**.

Let M be a model for a stochastic decision theory **SD**, then each branch instruction $q <?> a$ determines an element of the underlying Boolean algebra, which we may denote as $M(q <?> a)$. This element also determines a measure-preserving monic and its corresponding submodel, which we denote $M_{q <?> a}$. A *retrieval* of a conjunct C , in a model, is defined probabilistically as

$$Ret(C, M) = \frac{m(M(C))}{m(\top_M)},$$

where $M(C) = M(q_1 <?> a_1) \cap \dots \cap M(q_n <?> a_n)$.

These observations lead to the following results:

Theorem 12 Let t , t_1 , and t_2 be repeat-consistent stochastic decision terms and M be a non-trivial model, then

- (i) If M satisfies t and $Ret(C, t)$ exists, then $Ret(C, M) = Ret(C, t)$.
- (ii) If $t_1 \implies t_2$ and $Ret(C, t_2)$ exists, then $Ret(C, t_1)$ exists and $Ret(C, t_1) = Ret(C, t_2)$.
- (iii) If $Ret(C, t)$ does not exist, then there exist at least two models M_1 and M_2 both satisfying t for which $Ret(C, M_1) \neq Ret(C, M_2)$.

Proof:

The complete proof is given in Appendix B.5 on Page 176.

□

Definition 39

- (i) Let $\mathcal{C}$ be a set of conjuncts which includes $\top$, then C , a conjunct in $\mathcal{C}$, is said to be a **simple conjunct** in $\mathcal{C}$ if
 - (a) $C = \top \in \mathcal{C}$, or
 - (b) $C = q \langle ? \rangle a \ \& \ C'$, with C' being a simple conjunct in $\mathcal{C}$.
- (ii) A conjunct C is said to be **degenerate** in a stochastic term t (in a model M), if $Ret(C, t) = 0$ ($Ret(C, M) = 0$).

Lemma 12 Let

$$M(t) = \{M \mid M \text{ is a model which satisfies } t\} \text{ and}$$

$$\mathcal{C}_{M(t)} = \{C \mid \forall M_1, M_2 \in M(t), Ret(C, M_1) = Ret(C, M_2)\},$$

then C is non-degenerate in t , simple in $\mathcal{C}_{M(t)}$, and $C \& q \langle ? \rangle a \in \mathcal{C}_{M(t)}$ if and only if $Ret_s(q \langle ? \rangle a \mid C, t)$ exists.

Proof:

The proof is given in Appendix B.5 on Page 179.

□

Theorem 13 (Completeness)

$t_1 \implies t_2$ if and only if $M(t_1) \subseteq M(t_2)$.

Proof:

The proof is given in Appendix B.5 on Page 176.

□

4.7 SUMMARY AND CONCLUSION

In this chapter, we have described a mathematical framework for stochastic decision theory. We have shown, with respect to this framework, that the rules we have introduced for their structural manipulation are sound and complete, and they result in a consistent theory.

The applications of this theory range over a wide varieties of disciplines. Because probability calculations are simultaneously handled in the manipulations, in conformance with the laws of probability, classical problems of a stochastic nature, such as state estimation, can be approached and visualized via structural manipulations. In Chapter 7, a classical state estimation problem is re-cast from the stochastic decision theory point of view, and a cluster decomposition scheme of the terms of the theory is used to reduce the state estimation problem into solving a set of smaller state estimation sub-problems.

Another point of interest is the representation of uncertain knowledge for knowledge-based systems. In this area, the new capabilities provided by this theory for manipulating stochastic tree structures, retrieving information from it, and updating it, provide potential improvements to the current state of the art.

CHAPTER 5

PROBABILISTIC ENTAILMENT IN STOCHASTIC DECISION THEORY

As introduced in Chapter 4, stochastic decision theory can be used to represent uncertain knowledge. The tool of this representation is the stochastic term, which, within this theory and its identities, completely characterizes the knowledge. The theory, as it was shown in Chapter 4, not only can deal with complex dependence among the propositions (decisions) of the knowledge, but also with the case in which the conditional probabilities describing this dependence may be missing. This problem of incomplete information is a familiar occurrence in practically all realistic problems. In many instances, stochastic decision theory is incapable of recovering the needed information. Such instances occur mainly when the transposition rule is attempted or when a structural or general retrieval is needed. These problems were thoroughly discussed in Chapter 4 (see also [15]). This chapter addresses the following question: How can missing probabilities be estimated in an “optimal” way? This problem is often referred to in the literature as probabilistic entailment [55]. The consensus in approaching this problem is to complete the missing information, subjectively, in the sense of the maximum entropy principle. This principle has been traditionally used to determine unknown probability distributions possibly subject to some known constraints. This chapter solves the problem of probabilistic entailment using the maximum entropy principle in the presence of

1. constraints imposed by the stochastic decision theory, and

2. linear constraints.

The chapter also outlines algorithms for this completion in the context of decision theory.

Section 5.1 extends the concept of entropy, introduced in Subsection 3.3.2 of Chapter 3, to stochastic decision theory and gives a new recursive method to compute entropy for stochastic terms. Section 5.2 states the maximum entropy principle and its relevance to the probabilistic entailment problem and derives the uniformity results of the missing information in the presence of constraints imposed by the stochastic decision theory and outlines, in this case, a simple algorithm for probabilistic entailment. In the presence of linear constraints, the probabilistic entailment problem becomes equivalent to solving a set of nonlinear programs. Section 5.3 applies the results to an illustrative example. Section 5.4 concludes the chapter.

5.1 ENTROPY IN STOCHASTIC DECISION THEORY

Let t be a stochastic decision term of a stochastic decision theory as defined in Chapter 4 (see also [15,72]). The leaves of this stochastic term can be viewed as the possible events of the random experiment which consists of the combined sub-experiments of the decisions in this stochastic term. As a result, the entropy for this stochastic term is $H = -\sum_i p_i \log p_i$, with p_i being the probability of occurrence of the i^{th} leaf.

Lemma 13 *Let t be a stochastic term of a stochastic decision theory. The entropy in t is recursively computed as follows:*

- (i) If $t = q : [a_1/p_1 \gg t_1, \dots, a_n/p_n \gg t_n]$, then $H(t) = -\sum_{i=1}^n p_i \log p_i + \sum_{i=1}^n p_i H(t_i)$.
- (ii) If $t = x$, where x is a variable, then $H(t) = 0$.

Proof:

The proof is given in Appendix C.1 on Page 181. □

Lemma 14 *For two stochastic terms t_1 and t_2 , if $t_1 \implies t_2$, then $H(t_1) \geq H(t_2)$.*

Proof:

The proof is given in Appendix C.1 on Page 181. □

Corollary 2 *For two stochastic terms t_1 and t_2 , if $t_1 \iff t_2$, then $H(t_1) = H(t_2)$.*

Proof:

The proof is given in Appendix C.1 on Page 182. □

Lemmas 13, 14, and Corollary 13 render the problem of probabilistic entailment of stochastic terms, using the *maximum entropy principle*, trivial. This will be demonstrated in the next section.

5.2 THE MAXIMUM ENTROPY PRINCIPLE

Suppose we have of a stochastic term t which has incomplete information; some decisions do not appear in all the reachable paths of the term t . These decisions are not *reachable semi-essential*. A complete characterization of reachable semi-essential decisions is available in Chapter 4 (see also [15]). Suppose that a probability distribution for these decisions is required. This is the case, for instance, in which we seek a judgement on the independence of the decisions in a stochastic decision tree, as addressed in Chapter 6 (see [73]). This problem corresponds to the problems addressed by Cheeseman in [10] and Nilsson in [55]. There is a notable difference, however, that must be understood.

In stochastic decision theory, stochastic terms are always subject to the probabilistic constraints imposed by the models satisfying these terms. As a result, in stochastic decision theory, the unconstrained case corresponds to the constrained case addressed by Cheeseman and Nilsson with the constraints taking the form of (conditional) probability equations. Many other constraints (such as linear equalities or inequalities) cannot be viewed in the context of stochastic decision theory and when considered, constitute the constrained version of the problem.

In the presence of constraints (including the one imposed by stochastic decision theory), the unknown probabilities are constrained to a specified range. Nilsson [55] gave a geometric characterization of the range to which these probabilities are constrained. The problem of probabilistic entailment is to find unique values for these probabilities within the allowed range. These solution values are selected, according to the maximum entropy principle and are the best estimates, given the available information, of the true probability values in the sense that the least amount of the missing information is assumed. This solution is the probability distribution which maximizes entropy. This estimate can be computed using *Lagrange multipliers* [3]. In [10], Cheeseman outlined the steps which give a complete solution to the problem by solving $n + m$ Lagrange multiplier equations, where n is the number of the joint probabilities and m is the number of constraints imposed on these joint probabilities.

The method we proposed in Lemma 13 of computing the entropy in a stochastic term t avoids the computational complications by allowing the known part of the probability distribution to be separated from the unknown part and only deals only with the unknown part of the distribution.

5.2.1 The Unconstrained Case

In the absence of additional constraints, stochastic terms are always subject to the constraints imposed by the models satisfying these terms. As a result, from the stochastic decision theory point of view this is the unconstrained case of the problem. The stochastic decision theory constraints take the form of equations relating or prescribing some of (conditional) probabilities of the given stochastic term. In this case, the method of entropy computation proposed in Lemma 13 reduces the problem of probabilistic entailment to an assignment of uniform probability distribution on the missing information. The following theorem supports this claim.

Theorem 14 *All the conditional probability distributions of the missing information in a stochastic term t which achieve maximum entropy are uniform.*

Proof:

The complete proof is given in Appendix C.2 on Page 183.

□

The results of Theorem 14 reduces the probabilistic entailment to a simple assignment of uniform probability distribution on the paths of the missing stochastic terms. Note that, in view of Lemma 14, missing stochastic terms can be inserted at any height of the tree provided that the condition of “semi-essentiality” is respected. This can be achieved by completing the tree in appropriate paths. The next section gives an illustrative application example of these results.

5.2.2 The Constrained Case

In addition to the constraints imposed by stochastic decision theory, other constraints are possible. Linear or nonlinear equalities or inequalities are examples of such constraints.

As these constraints are commonly present, it is important to derive a general solution to the probabilistic entailment problem in the presence of such constraints. Only linear equality or inequality constraints are considered in this dissertation. Other constraints can be handled in similar fashion at the expense, perhaps, of some computation complexity. In the constrained case, the probabilistic entailment problem is equivalent to solving a set of nonlinear programs [46]; the nonlinearity is due to the objective functions of these programs, namely, the entropy functions. The following theorem summarizes this result.

Theorem 15 *All the conditional probability distributions of the missing information in a stochastic term t which achieve maximum entropy under linear constraints are solutions to a set of nonlinear programs.*

Proof:

The complete proof is given in Appendix C.2 on Page 184.

□

5.3 AN APPLICATION EXAMPLE

Consider a problem in which we seek to determine the position of a non-moving target in a four-quadrant grid (Figure 5.1). The states of the system can be selected to be the four numbered quadrants x_1, x_2, x_3 , and x_4 .

Suppose that our knowledge about the system is represented by the stochastic term t_1 of Figure 5.2 in which the probability distribution for the decision “top_or_bottom” is missing. Suppose that we want to evaluate the probability distribution over the states of the system given that the target is in a top quadrant, namely, $p(x_i \mid \text{top_or_bottom} \leq \text{top}), \forall i \in \{1, 2, 3, 4\}$.

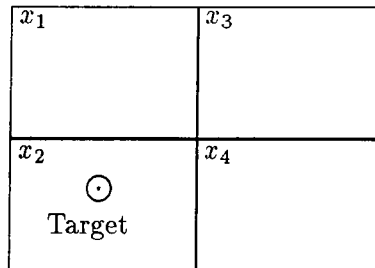


Figure 5.1: The 4-Quadrant Non-Moving Target Problem.

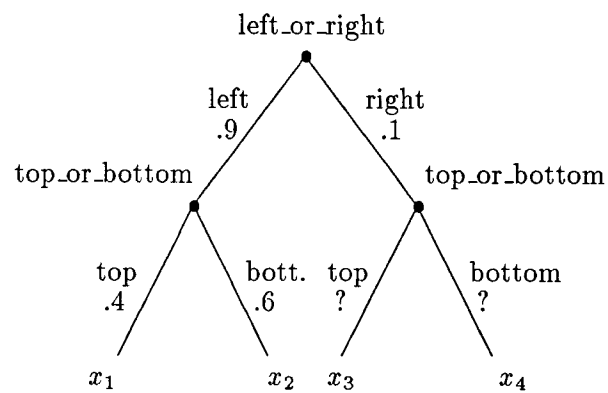


Figure 5.2: The Stochastic Term t_1 .

This conditional probability can be computed only if the decision “top_or_bottom” is semi-essential in the stochastic term t_1 , which it is, except the missing probability distribution is needed. According to the unconstrained case of the maximum entropy principle, these probabilities can be uniformly assigned. Figure 5.3 depicts the resulting stochastic term, t_2 .

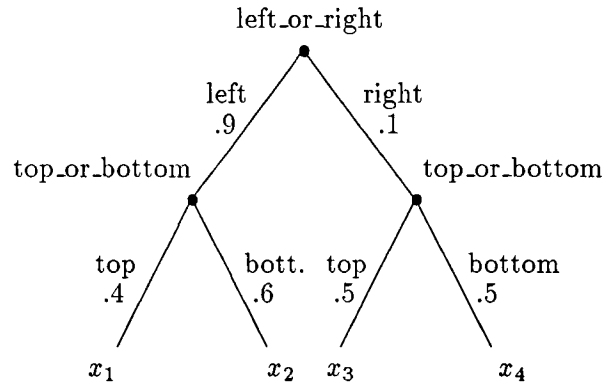


Figure 5.3: The Stochastic Term t_2 .

The desired conditional probability can be easily obtained by applying the inference rule [SD.4] to the stochastic term t_2 to put the decision “top_or_bottom” at the top of the stochastic term t_3 of Figure 5.4, from which we conclude that

$$\left\{ \begin{array}{l} p(x_1 \mid \text{top_or_bottom} \langle ? \rangle \text{top}) = 0.88, \\ p(x_2 \mid \text{top_or_bottom} \langle ? \rangle \text{top}) = 0.00, \\ p(x_3 \mid \text{top_or_bottom} \langle ? \rangle \text{top}) = 0.12, \\ p(x_4 \mid \text{top_or_bottom} \langle ? \rangle \text{top}) = 0.00. \end{array} \right.$$

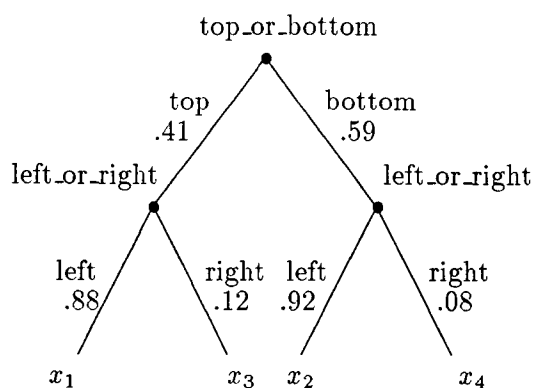


Figure 5.4: The Stochastic Term t_3 .

5.4 CONCLUDING REMARKS

In this chapter, we presented a recursive method for the computation of entropy in stochastic decision theory. The problem of probabilistic entailment in the context of the maximum entropy principle with and without constraints is solved. The no-constraint probabilistic entailment problem is reduced to a simple assignment of uniform probability distributions on the missing information. The constrained case (linear equalities or inequalities) is equivalent to solving a set of nonlinear programs.

The results of this chapter are very important to the stochastic decision theory. It allows many operations to be carried out, in a stochastic term, without excessive assumption on the model satisfying this stochastic term. In the following chapter, where the statistical independence relations among the decisions in a stochastic term is needed, the entailment results allow all such relations to be derived in an optimal way. This will contribute to the decomposition scheme in stochastic decision theory to be discussed in Chapter 7.

CHAPTER 6

STATISTICAL INDEPENDENCE IN STOCHASTIC DECISION THEORY

The stochastic decision theory setting depicts the concurrence of a possibly large number of theorems (or decisions) in a probabilistic space. Nominally, the theory assumes a complete dependence among these theorems. This assumption could severely undermine the efficiency and practicality of the stochastic decision theory representation. This fact has led many authors to assume, unrealistically, independence.

Many realistic problems exhibit “sparse dependence”, meaning that most of the theorems are actually statistically independent. This property may be the key to the representation and manipulation of realistic (large scale) problems. The question is: Given a stochastic term, how can one extract its dependence and independence classes in an efficient way, and can this lead to a useful decomposition of, for example, an estimation problem. This chapter addresses the first question while Chapter 7 addresses the second.

Section 6.1 re-casts the familiar concept of statistical independence in probability theory into the terminology of stochastic decision theory and gives the necessary and sufficient conditions for an arbitrary set of decisions, in a stochastic term, to be mutually statistically independent. Section 6.2 provides an algorithm which can be applied to a stochastic term to generate simultaneously all the dependence and independence classes of its decisions. Section 6.3 applies this result to a small example, and Section 6.4 gives some concluding remarks.

6.1 STATISTICAL INDEPENDENCE

Let t be a stochastic term in a stochastic decision theory. The set of measurable events in t is the set of paths in t of the form

$$P = \{q_1 \langle ? \rangle a_{1i}, q_2 \langle ? \rangle a_{2j}, \dots, q_m \langle ? \rangle a_{mk}\},$$

where the symbol $q_l \langle ? \rangle a_{li}$ represent the elementary event that query q_l is decided by response a_{li} belonging to its finite cardinality codomain set $\text{codom}(q_l)$.

Definition 40 *Two decisions q_1 and q_2 of the stochastic term t and with codomains of cardinalities m_1 and m_2 , respectively, are **statistically independent** if and only if*

$$\begin{aligned} p(q_1 \langle ? \rangle a_{1i}, q_2 \langle ? \rangle a_{2j}) &= p(q_1 \langle ? \rangle a_{1i})p(q_2 \langle ? \rangle a_{2j}), \\ \forall i \in \{1, \dots, m_1\} \text{ and } \forall j \in \{1, \dots, m_2\}. \end{aligned}$$

In the case of three decisions q_1, q_2 , and q_3 with codomain cardinalities m_1, m_2 , and m_3 , respectively, the independence conditions are

$$\left\{ \begin{array}{ll} p(q_1 \langle ? \rangle a_{1i}, q_2 \langle ? \rangle a_{2j}) &= p(q_1 \langle ? \rangle a_{1i})p(q_2 \langle ? \rangle a_{2j}) \\ p(q_1 \langle ? \rangle a_{1i}, q_3 \langle ? \rangle a_{3k}) &= p(q_1 \langle ? \rangle a_{1i})p(q_3 \langle ? \rangle a_{3k}) \\ p(q_2 \langle ? \rangle a_{2j}, q_3 \langle ? \rangle a_{3k}) &= p(q_2 \langle ? \rangle a_{2j})p(q_3 \langle ? \rangle a_{3k}) \\ p(q_1 \langle ? \rangle a_{1i}, q_2 \langle ? \rangle a_{2j}, q_3 \langle ? \rangle a_{3k}) &= p(q_1 \langle ? \rangle a_{1i})p(q_2 \langle ? \rangle a_{2j})p(q_3 \langle ? \rangle a_{3k}), \end{array} \right.$$

$$\forall i \in \{1, \dots, m_1\}, \forall j \in \{1, \dots, m_2\}, \text{ and } \forall k \in \{1, \dots, m_3\}.$$

This result can be extended to the case of l decisions $q_1, q_2, \dots, q_l$ of respective codomain cardinalities $m_1, m_2, \dots, m_l$ as follows. The l decisions $q_1, q_2, \dots, q_l$ are *mu-*

tu ally statistically independent if and only if

$$\begin{aligned}
& p(q_{l_1} \langle ? \rangle a_{l_1 i_1}, q_{l_2} \langle ? \rangle a_{l_2 i_2}, \dots, q_{l_{l_0}} \langle ? \rangle a_{l_{l_0} i_{l_0}}) \\
& = \\
& p(q_{l_1} \langle ? \rangle a_{l_1 i_1}) p(q_{l_2} \langle ? \rangle a_{l_2 i_2}) \cdots p(q_{l_{l_0}} \langle ? \rangle a_{l_{l_0} i_{l_0}}), \\
& \forall i_j \in \{1, \dots, m_{l_j}\}, \quad \forall j \in \{1, \dots, l_0\}, \quad \text{and} \quad \forall l_1, l_2, \dots, l_{l_0} \leq l, \\
& \text{with } l_\alpha \neq l_\beta \text{ whenever } \alpha \neq \beta.
\end{aligned}$$

These definitions are a straightforward application of Definition 25.

Theorem 16 *Let $q_1, q_2, \dots, q_l$ be l decisions with codomain cardinalities $m_1, m_2, \dots, m_l$, respectively. The following statements are equivalent:*

- (i) *The l decisions $q_1, q_2, \dots, q_l$ are mutually statistically independent.*
- (ii) *Any ordering $l_1, l_2, \dots, l_l$ of these decisions satisfies*

$$\begin{aligned}
& p(q_{l_2} \langle ? \rangle a_{l_2 i_2} \mid q_{l_1} \langle ? \rangle a_{l_1 i_1}) = p(q_{l_2} \langle ? \rangle a_{l_2 i_2}), \\
& p(q_{l_3} \langle ? \rangle a_{l_3 i_3} \mid q_{l_2} \langle ? \rangle a_{l_2 i_2}, q_{l_1} \langle ? \rangle a_{l_1 i_1}) = p(q_{l_3} \langle ? \rangle a_{l_3 i_3}), \\
& \quad \vdots \\
& p(q_{l_l} \langle ? \rangle a_{l_l i_l} \mid q_{l_{l-1}} \langle ? \rangle a_{l_{l-1} i_{l-1}}, \dots, q_{l_2} \langle ? \rangle a_{l_2 i_2}, q_{l_1} \langle ? \rangle a_{l_1 i_1}) = p(q_{l_l} \langle ? \rangle a_{l_l i_l}), \\
& \quad \forall i_j \in \{1, \dots, m_{l_j}\} \quad \text{and} \quad \forall j \in \{1, \dots, l\},
\end{aligned}$$

with m_{l_j} being the cardinality of the codomain of q_{l_j} .

- (iii) *There exists an ordering $l_1, l_2, \dots, l_l$ of these decisions which satisfies the above equations.*

Proof:

The proof is in Appendix D.1 on Page 186.

□

Proposition 16 is quite important in the sense that it provides a test for mutual statistical independence which can be performed on any stochastic term which depicts these decisions at its root in an arbitrary ordering. Note that in the independence test, there is no need to compute the *a priori* probabilities of the tree decisions, and all we need to check is that the conditional probabilities, in these independence conditions, remain constant. This greatly simplifies the independence test as the need for further stochastic term manipulations is avoided.

Corollary 3 *In order for the l decisions $q_{l_1}, q_{l_2}, \dots, q_{l_l}$ from a stochastic term t to be determined to be mutually statistically independent, they must all be reachable semi-essential in t .*

Proof:

The proof is given in Appendix D.1 on Page 188. □

This corollary allows non-reachable semi-essential decisions in stochastic terms to be disregarded for possible independence inter-relationships with other decisions. If there is not enough information, namely no prior probabilities, in the given stochastic term, the probabilistic entailment results derived in Chapter 5 can be used to complete the stochastic term at hand and make of it a full stochastic term in which every decision is reachable semi-essential.

6.2 DEPENDENCE/INDEPENDENCE GENERATING ALGORITHM

In this section, an algorithm which generates simultaneously all the dependence and independence inter-relationships among the reachable semi-essential decisions of a stochastic term t is outlined. Some preliminary developments are necessary.

6.2.1 Preliminaries

Let S be a set of N elements. The power set of S , denoted as $\mathcal{P}(S)$, is defined as the set of subsets of S and has exactly 2^N elements. A non empty element of a power set $\mathcal{P}(S)$ (a subset of S) is an *atom* if its only subsets are the empty set and itself.

Definition 41 Let S be a set, a an arbitrary element of S , and $\mathcal{P}(S)$ the power set of S . The a -fixed power set $\mathcal{P}_a(S)$ of the power set $\mathcal{P}(S)$ is defined by

$$\mathcal{P}_a(S) = \{X \in \mathcal{P}(S) \mid a \in X\}.$$

As an example, if $S = \{1, 2, 3\}$ and $a = 1$, then the power set of S is the set $\mathcal{P}(S) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$. The 1-fixed power set of $\mathcal{P}(S)$ is the set $\mathcal{P}_1(S) = \{\{1\}, \{1, 2\}, \{1, 3\}, \{1, 2, 3\}\}$.

Lemma 15 Let S be a set of l elements $a_1, a_2, \dots, a_l$, then $\mathcal{P}(S) = \mathcal{P}_{a_1}(S) \cup \mathcal{P}(S_1)$, where $S_1 = S - \{a_1\}$.

Proof:

The proof is given in Appendix D.2 on Page 188. □

Corollary 4 Let S be a set of l elements $a_1, a_2, \dots, a_l$, then

$$\mathcal{P}(S) = \left[\bigcup_{i=1}^{l-1} \mathcal{P}_{a_i}(S_{i-1}) \right] \cup \mathcal{P}(S_{l-1}),$$

where $S_0 = S$ and $S_i = S - \bigcup_{j=1}^i \{a_j\}$.

Proof:

Apply Lemma 15 repeatedly for the power sets $\mathcal{P}(S_{i-1})$, with $i \in \{1, 2, \dots, l-1\}$. □

Corollary 5 *Let S be a set of l elements $a_1, a_2, \dots, a_k, \dots, a_l$. For any $S_k = S - \cup_{j=1}^k \{a_j\}$, where $k \in \{1, 2, \dots, l-1\}$, we have that $\mathcal{P}_{a_k}(S_{k-1}) \cup \mathcal{P}(S_k) = \mathcal{P}(S_{k-1})$, with $S_0 = S$.*

Proof:

This is a direct result of Lemma 15. □

Corollary 5 provides a recursive formula to form power sets. Furthermore, it is important to point to the “nice” structure provided by this decomposition. This structure is best described by the following equivalence:

$$\forall k \in \{1, 2, \dots, l-1\}, X \in \mathcal{P}(S_k) \iff X \cup \{a_k\} \in \mathcal{P}_{a_k}(S_{k-1}).$$

6.2.2 Statistical Independence

At this point, we would like to make the connection between the above definitions and the concept of statistical independence. Let S be a set of l decisions $q_1, q_2, \dots, q_l$ from a stochastic decision theory. An **independence class** of S is any set of mutually statistically independent decisions. A **dependence class** of S is any set of statistically dependent (non-independent) decisions. An independence (dependence) class is said to be of order k , if it consists of k mutually statistically independent (statistically dependent) decisions. Clearly, every independence (dependence) class belongs to $\mathcal{P}(S)$. In fact, more can be said in the up-coming lemmas. The proofs are obvious, and are omitted.

Definition 42 *Let S be a set of l decisions $q_1, q_2, \dots, q_l$ from a stochastic decision theory. The power set of S , $\mathcal{P}(S)$, is the set of all the candidate independence (dependence) classes of these l decisions.*

Note that, in Definition 42, even though atoms of the set $\mathcal{P}(S)$ are irrelevant classes, they are considered to be independence classes. This will later appear convenient as the

structure of power sets is quite useful in the results to be derived.

Lemma 16 *Let S be a set of l decisions $q_1, q_2, \dots, q_l$ from a stochastic decision theory. If $D_0 \in \mathcal{P}(S)$ is an independence class, any $D_i \in \mathcal{P}(S)$ such that $D_i \subset D_0$ is also an independence class.*

This lemma excludes subsets of independence classes from further tests; they are automatically independence classes.

Lemma 17 *Let S be a set of l decisions $q_1, q_2, \dots, q_l$ from a stochastic decision theory. If $D_0 \in \mathcal{P}(S)$ is a dependence class, any $D_i \in \mathcal{P}(S)$ such that $D_i \supset D_0$ is also a dependence class.*

This lemma excludes sets which contain groups of dependent decisions from further testing; they are automatically dependence classes.

In view of the power set decomposition provided by Corollaries 4 and 5, the search for independence (dependence) classes is recursive. We outline it for the union $\mathcal{P}_{q_k}(S_{k-1}) \cup \mathcal{P}(S_k)$, for some $k \in \{1, 2, \dots, l-1\}$, given that the set of independence classes in $\mathcal{P}(S_k)$ has been determined. Using Lemma 17 and the structure of $\mathcal{P}_{q_k}(S_{k-1}) \cup \mathcal{P}(S_k)$, it is clear that only elements in $\mathcal{P}_{q_k}(S_{k-1})$ whose corresponding elements in $\mathcal{P}(S_k)$ are independence classes need to be checked. This is very crucial to the efficiency of the algorithm which follows. As to the dependence classes, they are generated simultaneously by the algorithm.

6.2.3 Generating Dependence/Independence Classes

Suppose we are given a stochastic term depicting a number of decisions, and we are asked to find all the independence classes. According to Corollary 3, if a decision is not reachable semi-essential in a stochastic term, it cannot be judged to be (even though it

might be) statistically independent from any other decision in the stochastic term. In this case, there is simply not enough information in the stochastic term to conclude about dependence or independence. There are two alternatives one can choose to deal with the problem:

- (i) Deal only with the stochastic term's reachable semi-essential decisions.
- (ii) Use the probabilistic entailment results presented in Chapter 5 to complete the stochastic term.

If one selects the first alternative, the stochastic term at hand must be manipulated into an equivalent stochastic term having all the reachable semi-essential decisions above the ones which are not and only keep the reachable semi-essential part. In the second alternative, all the decisions are reachable semi-essential. Thus, we can assume that we are given a stochastic term t depicting a set S of l reachable semi-essential decisions labeled as they are ordered in t by $q_1, q_2, \dots, q_i, \dots, q_l$.

An algorithm which generates the dependence/independence classes is outlined as follows:

Given that the set of independence classes $I(S_k)$ and the set of dependence classes $D(S_k)$ in $\mathcal{P}(S_k)$ have been determined ($I(S_k), D(S_k) \subseteq \mathcal{P}(S_k)$), do the following:

1. Form the test set $T_{q_k}(S_{k-1}) \subseteq \mathcal{P}_{q_k}(S_{k-1})$ as follows:

$$X \in I(S_k) \iff X \cup \{q_k\} \in T_{q_k}(S_{k-1}).$$

2. Test $T_{q_k}(S_{k-1})$ for independence.
3. After testing $T_{q_k}(S_{k-1}) = I_{q_k}(S_{k-1}) \cup D_{q_k}(S_{k-1})$, where $I_{q_k}(S_{k-1})$ is the set of independence classes in $T_{q_k}(S_{k-1})$ and $D_{q_k}(S_{k-1})$ is the set of dependence

classes in $T_{q_k}(S_{k-1})$. Thus, the set of independence classes $I(S_{k-1})$ in $\mathcal{P}(S_{k-1})$ is $I(S_{k-1}) = I(S_k) \cup I_{q_k}(S_{k-1})$ and the set of dependence classes $D(S_{k-1})$ in $\mathcal{P}(S_{k-1})$ is $D(S_{k-1}) = D(S_k) \cup D_{q_k}(S_{k-1})$.

4. Continue the same procedure with $I(S_{k-1})$ and $D(S_{k-1})$ until the set of independence classes $I(S_0)$ of $\mathcal{P}(S_0)$ and the set of dependence classes $D(S_0)$ of $\mathcal{P}(S_0)$, with $S_0 = S$, are determined.
5. The final set of independence classes I in $\mathcal{P}(S)$ is refined as follows:

$$I = I(S_0) - \{X \in I(S_0); \exists Y \in I(S_0), \text{ such that } X \subset Y\}.$$

The final set of dependence classes D in $\mathcal{P}(S)$ is refined as follows:

$$D = D(S_0) - \{X \in I(S_0); \exists Y \in I(S_0), \text{ such that } Y \subset X\}.$$

Testing independence in a test set $T(\cdot)$ is done as follows:

- (a) Pick an element X of $T(\cdot)$ with the highest cardinality.
- (b) Use the inference rules [SD.1]–[SD.5] to manipulate the stochastic term into an equivalent term in which the decisions in X are at its root.
- (c) Apply Theorem 16 to determine if the decisions are independent.
- (d) If X is an independence class, then, according to Lemma 16, every class $Y \in T(\cdot)$ such that $Y \subset X$ is automatically an independence class.
- (e) Set $T(\cdot)$ to $T(\cdot) - \{Y\}$ and go to (a).

6.3 EXAMPLE

Consider the binary stochastic term t_1 of Figure 6.1 which depicts three reachable semi-essential decisions q_1, q_2 , and q_3 . We would like to automatically uncover, from

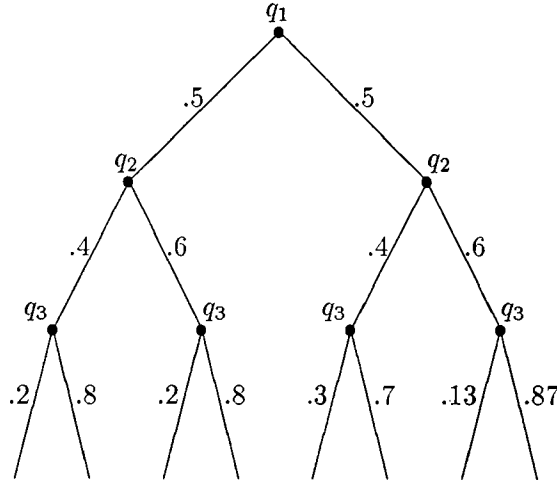


Figure 6.1: The Stochastic Term t_1 .

this tree structure, all the dependence/independence inter-relationships among the three featured decisions. Let $S = \{q_1, q_2, q_3\}$ be the set of these three decisions. The power set of S , $\mathcal{P}(S)$, is the set of all the candidate independence classes. Using the power set decomposition of Corollary 4, we write that

$$\mathcal{P}(S) = \mathcal{P}_{q_1}(\{q_1, q_2, q_3\}) \cup \mathcal{P}_{q_2}(\{q_2, q_3\}) \cup \mathcal{P}(\{q_3\}).$$

The dependence/independence generating algorithm will then proceed as follows:

- Set the initial conditions to $I(\{q_3\}) = \{\emptyset, \{q_3\}\}$ and $D(\{q_3\}) = \{\emptyset\}$. These are the set of independence classes and dependence classes in $\mathcal{P}(\{q_3\})$, respectively.

- $T(\{q_2, q_3\}) = \{\{q_2\}, \{q_2, q_3\}\}$. These are the only classes in $\mathcal{P}_{q_2}(\{q_2, q_3\})$ which need to be checked.
- Manipulate the stochastic term t_1 of Figure 6.1 into the stochastic term t_2 of Figure 6.2.
- From testing the stochastic term t_3 , the set $\{q_2, q_3\}$ is not an independence class.
- Set $I(\{q_2, q_3\}) = \{\emptyset, \{q_2\}, \{q_3\}\}$. This is the set of independence classes in $\mathcal{P}(\{q_2, q_3\})$. Set $D(\{q_2, q_3\}) = \{\{q_2, q_3\}\}$. This is the set of dependence classes in $\mathcal{P}(\{q_2, q_3\})$.
- $T(\{q_1, q_2, q_3\}) = \{\{q_1\}, \{q_1, q_2\}, \{q_1, q_3\}\}$. These are the only classes in $\mathcal{P}_{q_1}(\{q_1, q_2, q_3\})$ which need to be checked.
- Manipulate the stochastic term t_1 of Figure 6.1 into the stochastic term t_3 of Figure 6.3.
- From testing the stochastic term t_1 , the set $\{q_1, q_2\}$ is an independence class. Similarly, testing the stochastic term t_2 reveals that the set $\{q_1, q_3\}$ is also an independence class.
- Set $I(\{q_1, q_2, q_3\}) = \{\emptyset, \{q_1\}, \{q_2\}, \{q_3\}, \{q_1, q_2\}, \{q_1, q_3\}\}$. This is the set of independence classes in $\mathcal{P}(\{q_1, q_2, q_3\})$. Set $D(\{q_1, q_2, q_3\}) = \{\{q_2, q_3\}\}$. This is the set of dependence classes in $\mathcal{P}(\{q_1, q_2, q_3\})$.
- Finally, the set of independence classes is refined to $I = \{\{q_1, q_2\}, \{q_1, q_3\}\}$, while the set of dependent classes is $I = \{\{q_2, q_3\}\}$.

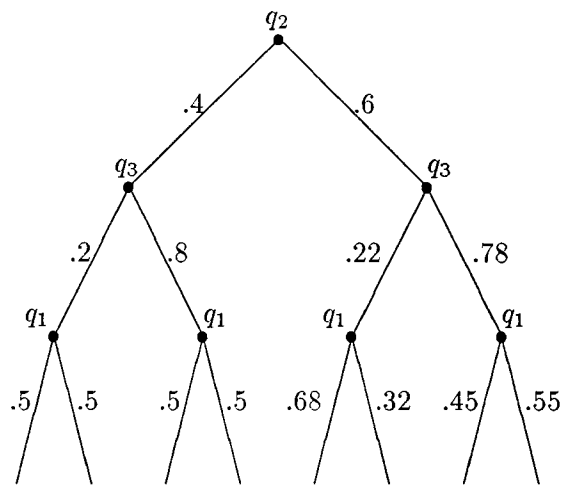


Figure 6.2: The Stochastic Term t_2 .

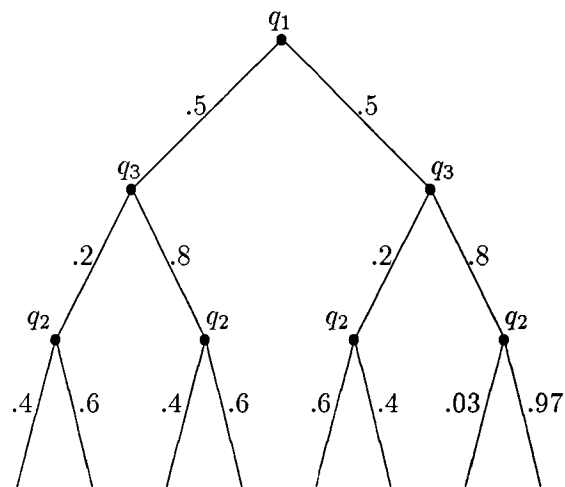


Figure 6.3: The Stochastic Term t_3 .

This algorithm produces a set of independence classes which consists of the elements of the set I and displays the following result:

1. The decisions q_1 and q_2 are statistically independent.
2. The decisions q_1 and q_3 are statistically independent.
3. The decisions q_2 and q_3 are statistically dependent.

6.4 CONCLUSION

This chapter describes an algorithm which can be performed on a stochastic term to determine all the dependence/independence inter-relationships among the reachable semi-essential decisions of this stochastic term. The question is how to take advantage of these results to alleviate the problem of state estimation's computational explosion in large scale systems. This problem is addressed in the following chapter where a cluster decomposition scheme for stochastic terms is proposed based on, first, the probabilistic entailment results derived in Chapter 5 and, second, the dependence/independence algorithm.

CHAPTER 7

A STATE ESTIMATION PROBLEM

In this chapter, we formulate a state estimation problem for a class of *static* finite state systems whose initial state probability distributions are described by a stochastic term. The classical solution is given, and the difficulties with large scale systems are demonstrated. The problem is then approached from the stochastic decision theory point of view. This provides a new insight into the problem of state estimation as the estimation procedure is viewed in terms of the basic stochastic term manipulations introduced in Chapter 4. This approach has also led to a cluster decomposition scheme for stochastic terms which optimally decomposes the problem of state estimation into a number of smaller sub-problems. This decomposition scheme, based on the probabilistic entailment results derived in Chapter 5 and also upon the statistical independence inter-relationships between the decisions of stochastic terms which have been uncovered in Chapter 6, is an attempt to alleviate the computational explosion of state estimation in large scale systems.

Section 7.1 formulates and solves a classical state estimation problem. Section 7.2 discusses the difficulties and limitations encountered in large scale systems. Section 7.3 approaches the problem of state estimation from the stochastic decision theory point of view and gives a simple illustration. Section 7.4 uses the results of probabilistic entailment derived in Chapter 5, the results of statistical independence discussed in Chapter 6, and an approximation method for joint probability distributions to propose a new cluster decomposition scheme for stochastic terms. Section 7.5 uses this cluster decomposition scheme to reduce the state estimation problem into a set of small sub-problems and illustrates the

results with an example. Section 7.6 summarizes these results and gives some concluding remarks.

7.1 A CLASSICAL STATE ESTIMATION PROBLEM

Consider a static finite state system [6] whose state is one value from a finite set $X = \{x_i \mid i = 1, N\}$ of possible states. Suppose that the *a priori* knowledge about the state of the system is represented by a stochastic term that consists of M decisions. We assume that there is a one-to-one correspondence between the set X of the system states and the leaves of the stochastic term describing the system; otherwise, minor modifications should be made. In our case, these decisions not only partition the state space into a number of disjoint sets, but also provide an initial evaluation of the *a priori* probability density functions of the system states, namely, $p(x_i)$, $\forall i = 1, \dots, N$, with $\sum_{i=1}^N p(x_i) = 1$.

Suppose that measurements were taken on a number (M') of observable decisions, which may or may not be represented in the stochastic term describing the initial knowledge about the system. Each measurement, on a given decision q_i , with arity m_i and codomain $\text{codom}(q_i) = \{a_{ij}, j = 1, \dots, m_i\}$, selects one element of that decision codomain. The set Y of all the possible observations is referred to as the *observation space*, and is

$$Y = \bigoplus_{i=1}^{M'} \text{codom}(q_i),$$

with the q_i 's being the observable decisions, from a stochastic decision theory **SD**, some of which are not necessarily featured in the stochastic term describing the system. We assume that this observation process provides *on-line* data about the system. These data must be used, in *real-time*, to update the current knowledge about the system. The observations are assumed to be corrupted by noise. We also assume that we are provided, *a priori*, with the conditional probability density function of each observation $y \in \text{codom}(q_j)$, for

some $j \in \{1, 2, \dots, M'\}$, conditioned over each of the system states, namely,

$$p(y \mid x_i), \quad \forall i = 1, \dots, N.$$

When an observation y is received, the estimate of the system state is updated to the *a posteriori* conditional probability density function of the states, conditioned over the received observation,

$$p(x_i \mid y), \quad \forall i = 1, \dots, N.$$

The state estimation problem [35,26] can now be formalized as follows: Given a sequence of observations,

$$Y_k = \{(y_1, q_{i_1}), (y_2, q_{i_2}), \dots, (y_k, q_{i_k})\},$$

with $y_k \in Y$ being the observation from an observable decision q_{i_k} (this means $y_k \in \text{codom}(q_{i_k})$), with $i_k \in \{1, \dots, M'\}$, and given that the *a posteriori* conditional probability density functions of the system states, based on Y_k ,

$$p(x_i \mid Y_k), \quad \forall i = 1, \dots, N,$$

is known, compute the updated *a posteriori* conditional probability density functions of the system states,

$$p(x_i \mid Y_{k+1}), \quad \forall i = 1, \dots, N,$$

where

$$Y_{k+1} = Y_k \cup \{y_{k+1}\} = \{y_1, y_2, \dots, y_k, y_{k+1}\},$$

when a new observation y_{k+1} ($y_{k+1} \in \text{codom}(q_{i_{k+1}})$) is received. For this observation, we are provided with

$$p(y_{k+1} \mid x_i), \quad \forall i = 1, \dots, N,$$

which is the state-conditioned probability density function of the new observation y_{k+1} .

The solution of this *recursive* filtering problem is the update

$$p(x_i | Y_{k+1}) = \frac{p(y_{k+1} | x_i) \times p(x_i | Y_k)}{\sum_{i=1}^N p(y_{k+1} | x_i) \times p(x_i | Y_k)}.$$

7.2 STATE ESTIMATION IN LARGE SCALE SYSTEMS

Some engineering problems, such as the ones dealing with power systems or communication networks, involve large scale discrete systems for which the number of states is quite large (in the thousands). For such systems, the amount of data inherent in the problem far exceeds the capacity of any foreseeable computer. In practice, limitations on the allowable complexity of the equipment are almost always imposed for operational and maintainability reasons. As a result, the performance of (the necessarily suboptimal) estimators for large scale systems is almost entirely determined by the size of the available machine and how the computer resources are used.

The state estimation procedure, as we described in the previous section, when applied to large scale systems, faces the serious problems of exponentially increasing (in the number of states) storage memory and computational operations which remain outside the bounds of engineering practicality despite the ever increasing performance of current computers.

For instance, consider a system described by a stochastic term t with n binary decisions (say $q_1, q_2, \dots, q_n$). This system may have up to 2^n states. The joint probability distribution defined by this stochastic term t , $p(q_1, q_2, \dots, q_n)$, must be encoded in the computer. The explicit storage of the joint probability distribution would require, in this particular case, a table with 2^n entries. The storage of this information as a stochastic term would need more machine memory (up to $2^{n+1} - 2$ entries) but has the advantage

of explicitly providing some marginal probabilities. For example, computing the marginal probability $p(q_i)$ from the joint probability would require summing $p(q_1, q_2, \dots, q_n)$ over all the 2^{n-1} combinations of the remaining $n - 1$ decisions, while in the stochastic term form, this probability is readily provided when decision q_i is at the root of the stochastic term t . The severity of these constraints upon the storage requirement worsens at each step of the estimation algorithm, as the probabilities of the observation process, conditioned upon every state of the system, must be encoded into the storage structure. This requires up to 2^n additional memory entries.

From the computational point of view, updating the *a priori* probabilities $p(x_i)$ at a given step of the estimation procedure after an observation y is recorded, is computing the *a posteriori* conditional probabilities $p(x_i | y)$, which involves dividing two marginal probabilities, each resulting from summation over an exponentially large number (in terms of the number of decisions) of combinations of decisions.

These requirements dictate a scheme in which the probability distributions are approximated by the use of simplifying assumptions, such as statistical independence. The next section approaches the state estimation problem from the stochastic decision theory point of view. In this context, a decomposition scheme is devised to reduce the storage problem and the computational explosion experienced with large scale systems.

7.3 THE STOCHASTIC DECISION THEORY APPROACH

7.3.1 The Insight of Stochastic Decision Theory

This estimation problem can be approached from the stochastic decision theory point of view by exploiting the structure of the stochastic term which describes the system. The estimation procedure can then be interpreted in terms of stochastic term manipulations.

To accomplish this, the observation process is incorporated into the stochastic term describing the system. This is done by the introduction, at each step of the estimation procedure and for each observation, of an additional decision, the *observation decision*, which depicts the observation conditional probabilities. This decision, which has a single branch, is appended to each leaf of the stochastic term describing the system before this observation is processed, with the corresponding conditional probabilities. The following theorem proves the above conjecture.

Theorem 17 *The recursive static state estimator is equivalent to transposing, to the root, the observation decision in the stochastic term which describes the system and its observation process before the current observation is recorded and to which this decision is appended.*

Proof:

The complete proof is given in Appendix E.1 on Page 190. □

Note that if the observations are perfectly received, the *retrieval properties* as introduced in Chapter 4 (see also [15]) can be used to generate an estimate of the state of the system. The above results are best illustrated by an example.

7.3.2 An Example

Consider a *static* system whose state is one value from a set $X = \{x_1, x_2, x_3, x_4\}$ of possible states. Suppose that the *a priori* knowledge about the state of the system is given by the stochastic term depicted in Figure 7.1 featuring three decisions q_1 , q_2 , and q_3 . This stochastic term makes available to us the *a priori* probability density functions of the four

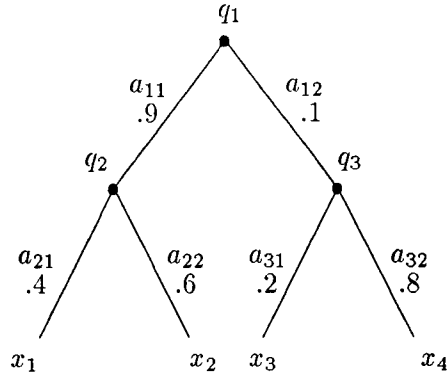


Figure 7.1: The Example's Stochastic Decision Tree.

states:

$$\left\{ \begin{array}{lcl} p(x_1) & = & .9 \times .4 = .36 \\ p(x_2) & = & .9 \times .6 = .54 \\ p(x_3) & = & .1 \times .2 = .02 \\ p(x_4) & = & .1 \times .8 = .08. \end{array} \right.$$

Suppose that the three decisions in the stochastic term of Figure 7.1 are observed. Each observation is associated with a particular decision q_i and selects an element from its codomain set $\text{codom}(q_i)$. Suppose that a first observation, y_1 , which monitors the decision q_2 of the stochastic term, is received, with y_1 being the selection of the branch labeled a_{21} , (*i.e.*, $y_1 = a_{21}$). Suppose that we are given the following conditional probability densities:

$$\left\{ \begin{array}{lcl} p(y_1 | x_1) & = & 0.6 \\ p(y_1 | x_2) & = & 0.3 \\ p(y_1 | x_3) & = & 0.7 \\ p(y_1 | x_4) & = & 0.2. \end{array} \right.$$

Given this, we can compute the *a posteriori* probability of the actual state being one of the four system states, given this observation and the *a priori* information,

$$\left\{ \begin{array}{lcl} p(x_1 | y_1) & = & \frac{p(y_1 | x_1) \times p(x_1)}{\sum_{i=1}^4 p(y_1 | x_i) \times p(x_i)} = .53 \\ p(x_2 | y_1) & = & \frac{p(y_1 | x_2) \times p(x_2)}{\sum_{i=1}^4 p(y_1 | x_i) \times p(x_i)} = .40 \\ p(x_3 | y_1) & = & \frac{p(y_1 | x_3) \times p(x_3)}{\sum_{i=1}^4 p(y_1 | x_i) \times p(x_i)} = .03 \\ p(x_4 | y_1) & = & \frac{p(y_1 | x_4) \times p(x_4)}{\sum_{i=1}^4 p(y_1 | x_i) \times p(x_i)} = .04. \end{array} \right.$$

This estimation procedure can be interpreted in terms of a transposition manipulation performed on the original stochastic term. The stochastic term is first augmented with the “observation decision”, by appending to each leaf x_i a decision $\mathcal{Y}$ consisting of a single branch labeled y_1 with the probability $p(y_1 | x_i)$. Figure 7.2 represents the augmented stochastic term of the example after observation y_1 has been received.

The transposition manipulation ([SD.4]) can be applied to move the observation decision $\mathcal{Y}$ to the top of the stochastic term. Figure 7.3 illustrates the manipulation. This produces the updated version of the stochastic term given in Figure 7.4. This updated version is obtained by omitting the observation decision and keeping the stochastic term probability distribution conditioned upon this observation. This stochastic term tells us that

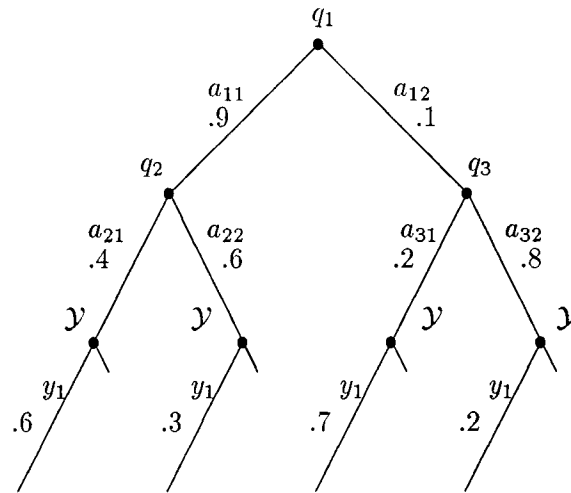


Figure 7.2: The Augmented Stochastic Decision Tree.

$$\left\{ \begin{array}{l} p(x_1 | y_1) = .93 \times .57 = .53 \\ p(x_2 | y_1) = .93 \times .43 = .40 \\ p(x_3 | y_1) = .07 \times .49 = .03 \\ p(x_4 | y_1) = .07 \times .51 = .04. \end{array} \right.$$

These results agree with the ones obtained by direct computation. These *a posteriori* conditional probability density functions are the estimates of the system states based upon the observation y_1 . These will become the *a priori* conditional probability density of the states for the next stage of the estimation process, when a second observation is received. This means that the stochastic term of Figure 7.4 will be used for the next stage transposition manipulation.

Suppose now that the observations are perfectly received and that definite arguments are recorded for the decisions. In this case, the retrieval procedure provides the estimate of

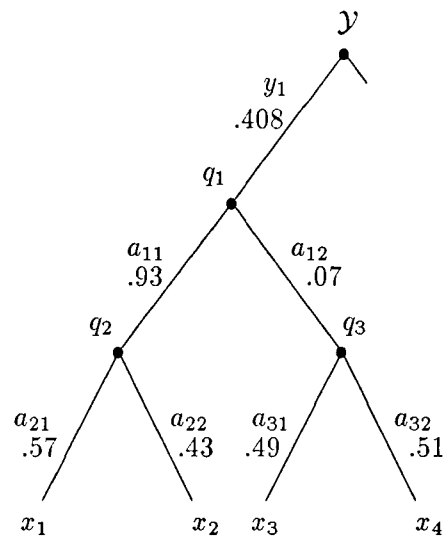


Figure 7.3: The Augmented Decision Tree after the Estimation Procedure.

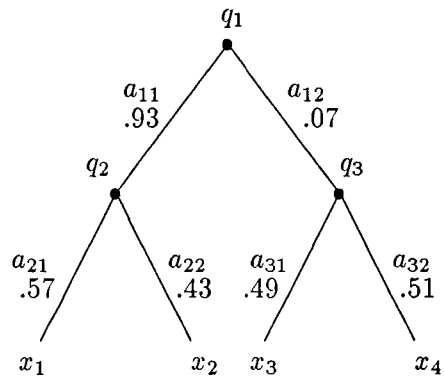


Figure 7.4: The Updated Stochastic Decision Tree of the Example.

the system state. Referring to Figure 7.1, suppose that our observation y_1 , which selects the branch a_{12} of decision q_1 , was recorded with certainty. Let C be the conjunction $q_1 <?> a_{12}$, which means that branch a_{12} is selected for decision q_1 . Two queries, Q_1 and Q_2 (see Chapter 4 or [15]), can be conditioned upon this conjunction. They are

$$\begin{cases} Q_1 &= p(q_2 <?> a_{21} \mid q_1 <?> a_{12}) \\ Q_2 &= p(q_2 <?> a_{22} \mid q_1 <?> a_{12}). \end{cases}$$

The stochastic term of Figure 7.1 already has the required form described in Chapter 4 (also in [15]) for these two queries, and therefore no manipulation is needed. This stochastic term tells us that both Q_1 and Q_2 can be retrieved in that stochastic term and that their retrievals are

$$\begin{cases} Q_1 &= p(q_2 <?> a_{21} \mid q_1 <?> a_{12}) = .20 \\ Q_2 &= p(q_2 <?> a_{22} \mid q_1 <?> a_{12}) = .80. \end{cases}$$

Thus, the *a posteriori* probability density functions of the system states are

$$\begin{cases} p(x_1 \mid y_1) &= .00 \\ p(x_2 \mid y_1) &= .00 \\ p(x_3 \mid y_1) &= .20 \\ p(x_4 \mid y_1) &= .80. \end{cases}$$

7.4 A CLUSTER DECOMPOSITION SCHEME

In this section, a cluster decomposition scheme for stochastic terms is proposed. The motivation behind this decomposition stems from the difficulties of storage and computation explosion encountered with the state estimation problem in large scale systems. The proposed decomposition scheme relies upon a method which optimally approximates the joint probability distribution in a given stochastic term by another joint probability

distribution which can be represented via a set of smaller (in size) stochastic terms. These smaller stochastic terms optimally approximate the information contained in the original stochastic term and the independence/dependence classes of this term. This stochastic term cluster decomposition then allows the state estimation problem to be reduced into a set of smaller sub-problems, reducing the storage and computation required for a large scale system.

7.4.1 Objective

As we recall in Chapter 6, we are motivated from the fact that most realistic problems exhibit the property of sparse dependence, meaning that in a given stochastic term, most decisions are mutually statistically independent from one another. The decomposition we propose in this chapter exploits this property. Ideally, one would like to replace a given stochastic term t whose decisions belong to a set S by a set of equivalent stochastic terms t_i , each of which has decisions belonging to a set S_i , with the S_i 's forming a partition of the set S . This partition $\{S_i\}$ is further required to have the following properties:

- (i) The decisions in S_i are statistically dependent upon one another.
- (ii) Every decision in S_i is statistically independent from any other decision in S_j , $\forall j \neq i$.

Such a decomposition is, unfortunately, not always possible because the relation “*is dependent upon*” is not an equivalence relation. This relation is only symmetric and it may be considered reflexive. One can, however, artificially make this relation an equivalence relation by enclosing into the dependence classes the decisions which fail transitivity. The price for this is that, in a given cluster, not all the decisions are dependent upon one another. The important point of pairwise independence between clusters, however, still holds after this operation.

Making the relation “*is dependent upon*” reflexive is a matter of choice. In chapter 6, the relation “*is independent of*” was considered reflexive for convenience. Considering the relation “*is dependent upon*” reflexive will not contradict this in any way. It remains to make it transitive. This is done by completing the classes for which transitivity did not work. Formally, this is done as follows:

Call the binary relation “*is dependent upon*” $\mathcal{R}$ and consider the binary relation $\mathcal{R}_0$ defined on the set S of decisions by

$$q_1 \mathcal{R}_0 q_2 \iff \begin{cases} q_1 \mathcal{R} q_2, \text{ or} \\ \exists \text{ a sequence of decisions in } S, \{q_a, q_b, \dots, q_n\}, \\ \text{such that } q_1 \mathcal{R} q_a, q_a \mathcal{R} q_b, \dots, q_n \mathcal{R} q_2, \end{cases}$$

for any $q_1, q_2 \in S$.

Lemma 18 *The binary relation $\mathcal{R}_0$ is an equivalence relation.*

Proof:

The complete proof is given in Appendix E.2 on Page 191. □

As the relation $\mathcal{R}_0$ is an equivalence relation on the set of decisions S , its equivalence classes define a unique partition on this set. This partition precisely determines the desired cluster decomposition.

7.4.2 Probabilistic Entailment

As cluster decomposition is based upon statistical independence, one needs to first complete the stochastic term at hand in order to uncover all the dependence/independence inter-relationships between the decisions of the term. In Chapter 5, we introduced an optimal way of completing stochastic terms in the context of the maximum entropy principle. In terms of the decomposition we are seeking, this completion means that the least

additional information have been assumed and that this decomposition is optimally approximative of the original stochastic term. From now on, we assume that the stochastic term at hand is completed (full).

7.4.3 Statistical Dependence

The algorithm described in Chapter 6 generates simultaneously all the independence and dependence classes in a full stochastic term. Let D be the set of all dependent classes of decisions in a given stochastic term t . In Subsection 7.4.1 the binary relation $\mathcal{R}_0$ which we defined from the binary relation $\mathcal{R}$, “*is dependent upon*”, was proven to be an equivalence relation. The equivalence classes of this relation form a partition of the set of decisions S ; call this partition set $\mathcal{P}$. Let D_f be the set obtained from the set D by repeatedly fusing together all the non-disjoint elements of the set D . The set D_f can be formally described as follows:

$$X \in D_f \iff \begin{cases} X \in D, \text{ or} \\ X = \cup_{i=1}^n X_i, \text{ with } X_i \in D, \text{ and } \forall i, \exists j \mid X_i \cap X_j \neq \emptyset. \end{cases}$$

Clearly, the set D_f defines a partition on the set of decisions in S . The following lemma claims the following:

Lemma 19 $\mathcal{P} = D_f$.

Proof:

The proof is given in Appendix E.2 on Page 192. □

Lemma 19 provides a method for determining the partition defined by the binary relation $\mathcal{R}_0$, simply fuse repeatedly together all the non-disjoint dependent classes of the set D . This determines the desired partition of the decisions of the stochastic term t . In view of this partition, the stochastic term t is manipulated, by transposition, into the

equivalent stochastic term t' for which decisions of the same partition are gathered together to form a subterm. The decomposition we wish to have will substitute for the stochastic term t the subterms taken independently from one another. This is, unfortunately, not possible because the conditional probabilities of these subterms may vary from instance to instance in the original stochastic term. Such decomposition is only possible when these partitioning sets are all singleton sets, meaning that all the decisions in t are mutually statistically independent. In realistic problems, though, this is very unlikely to happen; when it does, a decomposition into the prior probabilities is possible. This case is, however, uninteresting from the stochastic decision theory point of view as this theory is mainly devised to deal with complex dependencies. To illustrate these points, let's take the example of a stochastic term t in n decisions and suppose that its set D_f contains two classes: D_1 containing n_1 decisions and D_2 containing n_2 decisions ($n = n_1 + n_2$). If $D_f = \{D_1, D_2\}$, then there exists a stochastic term, t' , equivalent to the stochastic term t such that

$$t' = W_1 : [a_{11}/p_{11} \gg W_2 : [a_{21}/p_{21}^1 \gg x_{11}, \dots, a_{2n_2}/p_{2n_2}^1 \gg x_{1n_2}], \dots \\ \dots, a_{1n_1}/p_{1n_1} \gg W_2 : [a_{21}/p_{21}^{n_1} \gg x_{n_11}, \dots, a_{2n_2}/p_{2n_2}^{n_1} \gg x_{n_1n_2}]] ,$$

with W_1 consisting of only the n_1 decisions of D_1 , and W_2 consisting of only the n_2 decisions of D_2 . Contrary to one's expectation, the W_2 's may carry different conditional probabilities. Examples of when this occurs can be easily constructed (see for instance the example of the stochastic term t of Figure 7.5 on Page 142). This is certainly a major inconvenience to cluster decomposition. This difficulty can be avoided if one approximates the conditional probabilities of the subterm W_2 by its prior probabilities. In other words, we assume that

$$p_{2i}^j \approx p_{2i}, \quad \forall i = 1, \dots, n_2 \quad \text{and} \quad \forall j = 1, \dots, n_1,$$

where the p_{2i} 's are the probability obtained for the subterm W_2 when it is transposed to the root of the stochastic term t' . The stochastic term t can then be decomposed (approximately) into two stochastic term t_1 and t_2 such that

$$t_1 = W_1 : [a_{11}/p_{11} \gg (x_{11} \cup \dots \cup x_{1n_2}), \dots, a_{1n_1}/p_{1n_1} \gg (x_{n_1 1} \cup \dots \cup x_{n_1 n_2})]$$

and

$$t_2 = W_2 : [a_{21}/p_{21} \gg (x_{11} \cup \dots \cup x_{n_1 1}), \dots, a_{2n_2}/p_{2n_2} \gg (x_{1n_2} \cup \dots \cup x_{n_1 n_2})],$$

in which the variables have been properly unified.

This raises the important questions of whether or not this approximation is reasonable, how close it is from the original stochastic term, and whether a better approximation exists. The following subsection justifies this approximation and shows that, under certain conditions, it is optimal in the sense of the maximum entropy principle.

7.4.4 The Approximation Method

The problem of approximating a stochastic term by another stochastic term of lower height reduces to the problem of approximating a discrete joint probability distribution of a given order by a probability distribution of lower order. This is because a stochastic term is in fact no other than a convenient tool for writing joint probability distributions by use of the Chain rule. For example, for the two binary variables x_1 and x_2 , there are two ways of writing the joint probability distribution $p(x_1, x_2)$:

$$\left\{ \begin{array}{lcl} p(x_1, x_2) & = & p(x_2 | x_1)p(x_1) \\ p(x_1, \bar{x}_2) & = & p(\bar{x}_2 | x_1)p(x_1) \\ p(\bar{x}_1, x_2) & = & p(x_2 | \bar{x}_1)p(\bar{x}_1) \\ p(\bar{x}_1, \bar{x}_2) & = & p(\bar{x}_2 | \bar{x}_1)p(\bar{x}_1), \end{array} \right.$$

or

$$\left\{ \begin{array}{l} p(x_1, x_2) = p(x_1 | x_2)p(x_2) \\ p(\bar{x}_1, x_2) = p(\bar{x}_1 | x_2)p(x_2) \\ p(x_1, \bar{x}_2) = p(x_1 | \bar{x}_2)p(\bar{x}_2) \\ p(\bar{x}_1, \bar{x}_2) = p(\bar{x}_1 | \bar{x}_2)p(\bar{x}_2). \end{array} \right.$$

The first set of equations corresponds to the stochastic term t_1 , with

$$t_1 = x_1 : [x_1/p(x_1) \gg x_2 : [x_2/p(x_2 | x_1) \gg \Omega, \bar{x}_2/p(\bar{x}_2 | x_1) \gg \Omega], \\ \bar{x}_1/p(\bar{x}_1) \gg x_2 : [x_2/p(x_2 | \bar{x}_1) \gg \Omega, \bar{x}_2/p(\bar{x}_2 | \bar{x}_1) \gg \Omega]],$$

while the second set of equations corresponds to the stochastic term t_2 , with

$$t_2 = x_2 : [x_2/p(x_2) \gg x_1 : [x_1/p(x_1 | x_2) \gg \Omega, \bar{x}_1/p(\bar{x}_1 | x_2) \gg \Omega], \\ \bar{x}_2/p(\bar{x}_2) \gg x_1 : [x_1/p(x_1 | \bar{x}_2) \gg \Omega, \bar{x}_1/p(\bar{x}_1 | \bar{x}_2) \gg \Omega]].$$

Note that t_1 and t_2 are transposition equivalent.

The problem of approximating a joint probability distribution of a given order by a probability distribution which is a function of some of its lower order distributions has been addressed by many authors [8,14,31,41,44]. The work of Hartmanis in [31] was motivated by storage problems from information and coding theory. Hartmanis defined, for a given probability distribution, the set of probability distributions of higher order which are consistent with this distribution and which he called *extensions* of the original distribution. For a given probability distribution, there are infinitely many extensions. This definition is closely related to the concepts of *marginal probability distributions* and *lower order probability distributions*. Given a set of variables $X = \{x_1, x_2, \dots, x_N\}$ a probability distribution p on X , a probability distribution p_m on a subset X_m of X is said to be a **marginal probability distribution** of p if p_m results from p when the latter is properly summed [56]. A probability distribution p_l is said to be a **lower order probability distribution** if it involves only variables from a proper subset X_l of X .

Clearly, every marginal probability distribution of a given probability distribution is also a lower order distribution of that given distribution, but the converse is not true, in general. An equivalent definition to that used by Hartmanis [31] for extensions, but without assuming ergodicity, is

Definition 43 *Given a set of variables $X = \{x_1, x_2, \dots, x_N\}$ and a probability distribution p on X , then a probability distribution p_e on a subset X_e of X is said to be an **extension** of the probability distribution p if p_e is a marginal probability distribution of p .*

There is a simple method, inspired from stochastic decision theory, for constructing extensions of lower order distributions. Given a lower order distribution, an extension can be obtained by appending to the leaves of the stochastic term corresponding to the given lower distribution the desired subterm whose conditional probabilities sum to 1.

In [44], Lewis argued that if this approximation is to alleviate the storage problem, we should limit ourselves to approximations which use only component distributions that are particularly important in characterizing the distribution to be approximated. However, the problem of picking up the most influential components in a distribution is ill-posed as no definite meaning for the word “influential” can be made. It all depends on how much information one is willing to throw away (from the original distribution) to solve a particular problem. For that reason, a measure of the quality of the approximation and thus a selection criterion has to be established. This criterion, based on the concept of entropy, is an information theoretic measure of how close an approximating probability distribution is to the actual probability distribution. This measure is defined as follows:

Definition 44 *Given an n^{th} order probability distribution $p(x_1, x_2, \dots, x_n)$ with probabilities $p_1, p_2, \dots, p_m$, and another n^{th} order probability distribution $p'(x_1, x_2, \dots, x_n)$ with*

probabilities $p'_1, p'_2, \dots, p'_m$, the quantity

$$I = \sum_{i=1}^m p_i \log \frac{p_i}{p'_i}$$

is a measure of how close the two distributions are from each other.

This measure can be interpreted as the difference between the information contained in p and that contained in p' about p . More importantly, it has the following interesting property: $I \geq 0$, with $I = 0$ if and only if the two probability distributions are identical.

Lewis' formulation of the approximation problem assumes that several of the lower order distributions are available or known and that the approximating candidate distributions are the common extensions of these lower order distributions. A particularly tractable solution to this approximation problem is the product approximation. The **product approximation** of a given probability distribution, as defined by Lewis [44], is an approximation to this probability distribution made up of the product of some of its lower order probability distributions such that the product is a common extension of these lower order components. Intuitively, this corresponds to the constraint that these lower order distributions must be picked up from a single Chain rule writing of the higher order distribution. This approach is certainly not the general case, but it perfectly fits, as it will later become clear, our objective of stochastic term decomposition. Lewis then derived the following important result:

Theorem 18 (Lewis' theorem) *Given a probability distribution P and a set of some of its lower order probability distributions $P_1, P_2, \dots, P_m$ such that their product $P' = P_1 P_2 \dots P_m$ is a product approximation to P , then the distribution P' , of all the possible extensions of $P_1, P_2, \dots, P_m$, has maximum entropy.*

Proof:

The proof, due to Lewis [44], is given in Appendix E.2 on Page 193. □

Theorem 18 is very important as it gives a quite simple way of obtaining the optimal approximation for a given probability distribution provided that a suitable set of lower order distributions are available. The optimization is in the sense of maximum entropy which we adopted in Chapter 5. Lewis then uses the closeness measure of probability distribution to compare between various product approximations and concludes that better product approximations are obtained if for two interchangeable components, the component with no independence is used. The product approximation's measure of closeness to the actual distribution takes a particularly simple form. For instance, if the probability distribution $P' = P_1 P_2 \cdots P_m$ is an approximation of the distribution P , we have

$$\begin{aligned}
 I &= \sum_{i=1}^m p_i \log \frac{p_i}{p'_i} \\
 &= \sum_{i=1}^m p_i \log \frac{p_i}{p_{1i} p_{2i} \cdots p_{mi}} \\
 &= \sum_{i=1}^m (p_i \log p_i - p_i \log p_{1i} - p_i \log p_{2i} - \cdots - p_i \log p_{mi}) \\
 &= \sum_{i=1}^m (p_i \log p_i - p_{1i} \log p_{1i} - p_{2i} \log p_{2i} - \cdots - p_{mi} \log p_{mi}) \\
 &= H(P) - (H(P_1) + H(P_2) + \cdots + H(P_m)).
 \end{aligned}$$

This means that a measure of closeness of a distribution to its product approximation is the difference between the entropy of the original distribution and the sum of the entropies of each of its approximating products.

As we recall in Subsection 7.4.3, a decomposition for stochastic terms was proposed. Clearly, the probability distribution defined by this decomposition determines a set of lower order distributions. The product of these lower order distribution is a product approximation. This product corresponds to the decomposition of a stochastic term into

a set of lower height stochastic terms taken separately. As a result, the decomposition we proposed is optimal in the sense of the maximum entropy principle.

7.5 THE CLUSTER DECOMPOSITION AND THE STATE ESTIMATION PROBLEM

This section demonstrates how the cluster decomposition scheme we devised for stochastic terms can be used to decompose the problem of state estimation into a set of smaller sub-problems. An illustrative example is also given.

7.5.1 Decomposition of the State estimation Problem

Consider a full stochastic term t with n decisions describing a M -state system. Suppose that, at a given stage an observation y is received along with the conditional probabilities $p(y | x_i)$, $\forall i = 1, 2, \dots, N$. In order to update $p(x_i)$ to $p(x_i | y)$, $\forall i = 1, 2, \dots, M$, the stochastic term t is augmented with the observation decision and the state-conditioned probability. Call this augmented term t' . The application of the cluster decomposition method to the stochastic term t' yields a family of stochastic terms $\{w_i, i = 1, 2, \dots, m\}$ of lower height. The observation decision appears in only one of these subterms, say w_{i_0} . Let $\{q_{ij}, j = 1, 2, \dots, j_i\}$ be the family of decisions in w_i . Recall that for any i , with $i = 1, 2, \dots, m$, the decision q_{ij} is statistically independent from $q_{i'j'}$, for any $i' \neq i$ and $j' = 1, 2, \dots, j_{i'}$. The decomposition method assumes that the clusters are approximately statistically independent. The following development gives the "decomposed" solution of

the state estimation problem:

$$\begin{aligned}
& p(q_{11}, \dots, q_{1j_1}, \dots, q_{i_0 1}, \dots, q_{i_0(k-1)}, y, q_{i_0(k+1)}, \dots, q_{i_0 j_{i_0}}, \dots, q_{m1}, \dots, q_{mj_m}) \\
& = \\
& p(q_{11}, \dots, q_{1j_1}) \cdots p(q_{i_0 1}, \dots, q_{i_0(k-1)}, y, q_{i_0(k+1)}, \dots, q_{i_0 j_{i_0}}) \cdots p(q_{m1}, \dots, q_{mj_m}).
\end{aligned}$$

As a result, the *posteriori* probability

$$\begin{aligned}
& p(q_{11}, \dots, q_{1j_1}, \dots, q_{i_0 1}, \dots, q_{i_0(k-1)}, q_{i_0(k+1)}, \dots, q_{i_0 j_{i_0}}, \dots, q_{m1}, \dots, q_{mj_m} \mid y) \\
& = \\
& \frac{p(q_{11}, \dots, q_{1j_1}, \dots, q_{i_0 1}, \dots, q_{i_0(k-1)}, y, q_{i_0(k+1)}, \dots, q_{i_0 j_{i_0}}, \dots, q_{m1}, \dots, q_{mj_m})}{p(y)} \\
& = \\
& \frac{p(q_{11}, \dots, q_{1j_1}) \cdots p(q_{i_0 1}, \dots, q_{i_0(k-1)}, y, q_{i_0(k+1)}, \dots, q_{i_0 j_{i_0}}) \cdots p(q_{m1}, \dots, q_{mj_m})}{p(y)} \\
& = \\
& p(q_{11}, \dots, q_{1j_1}) \cdots p(q_{i_0 1}, \dots, q_{i_0(k-1)}, q_{i_0(k+1)}, \dots, q_{i_0 j_{i_0}} \mid y) \cdots p(q_{m1}, \dots, q_{mj_m}).
\end{aligned}$$

From above, the procedure of state estimation reduces to locally updating the subterm containing the observation decision and then taking the product of the corresponding probabilities for the update. The next subsection illustrates the procedure by an example.

7.5.2 An Example

Consider the stochastic term t of Figure 7.5 already augmented with the observation decision $\mathcal{Y}$. The technique we described in Chapter 6 can be used to uncover all the independence and dependent classes. The following results are obtained:

1. The decisions q_2 and $\mathcal{Y}$ are statistically dependent.
2. The decisions q_1 and q_2 are mutually statistically independent.
3. The decisions q_1 and $\mathcal{Y}$ are mutually statistically independent.

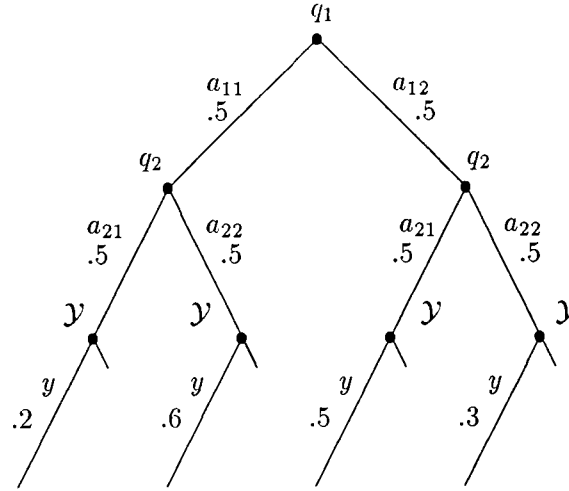


Figure 7.5: The Already Augmented Stochastic Term.

The cluster decomposition of t leads, in view of these independence/dependence classes, to two stochastic terms, t_1 and t_2 , depicted in Figure 7.6. How close this decomposition is to the original stochastic term is given by the measure of closeness of the two corresponding probability distributions. In this case this measure is

$$I = H(t) - H(t_1) - H(t_2) = 0.18.$$

This corresponds to an absolute error ($I/H(t)$) of less than 10%.

Transposing the observation decision, only in the term where it appears (t_2), leads to the stochastic terms of Figure 7.7 which gives that

$$\left\{ \begin{array}{l} p(q_1 \langle ? \rangle a_{11}, q_2 \langle ? \rangle a_{21} \mid y) = p(q_2 \langle ? \rangle a_{21} \mid y)p(q_1 \langle ? \rangle a_{11}) = 0.22 \\ p(q_1 \langle ? \rangle a_{11}, q_2 \langle ? \rangle a_{22} \mid y) = p(q_2 \langle ? \rangle a_{22} \mid y)p(q_1 \langle ? \rangle a_{11}) = 0.28 \\ p(q_1 \langle ? \rangle a_{12}, q_2 \langle ? \rangle a_{21} \mid y) = p(q_2 \langle ? \rangle a_{21} \mid y)p(q_1 \langle ? \rangle a_{12}) = 0.22 \\ p(q_1 \langle ? \rangle a_{12}, q_2 \langle ? \rangle a_{22} \mid y) = p(q_2 \langle ? \rangle a_{22} \mid y)p(q_1 \langle ? \rangle a_{12}) = 0.28. \end{array} \right.$$

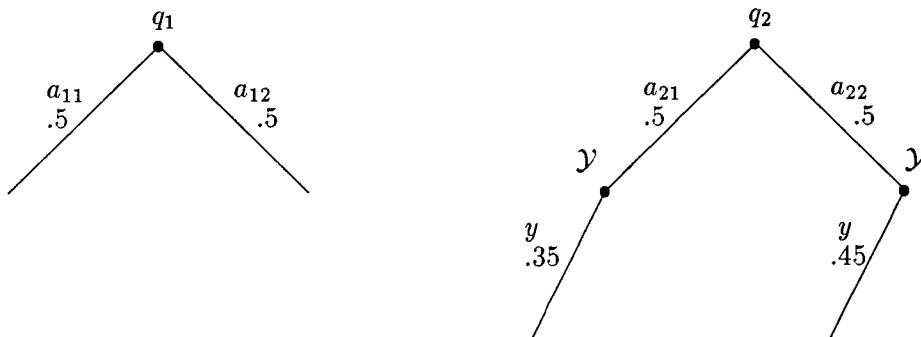


Figure 7.6: The Terms of the Cluster Decomposition.

These approximate updates are to be compared with the exact solutions which are obtained by transposing, in the stochastic term of Figure 7.5, the observation decision $\mathcal{Y}$ to the top. They are

$$\left\{ \begin{array}{l} p(q_1 <?> a_{11}, q_2 <?> a_{21} \mid y) = 0.125 \\ p(q_1 <?> a_{11}, q_2 <?> a_{22} \mid y) = 0.375 \\ p(q_1 <?> a_{12}, q_2 <?> a_{21} \mid y) = 0.3125 \\ p(q_1 <?> a_{12}, q_2 <?> a_{22} \mid y) = 0.1875. \end{array} \right.$$

7.6 CONCLUDING REMARKS

This chapter has approached the state estimation problem from the stochastic decision theory point of view. This approach not only visualizes the state estimation procedure as stochastic term manipulations, but also allows a cluster decomposition scheme to be devised, reducing the state estimation problem into a set of smaller problems.

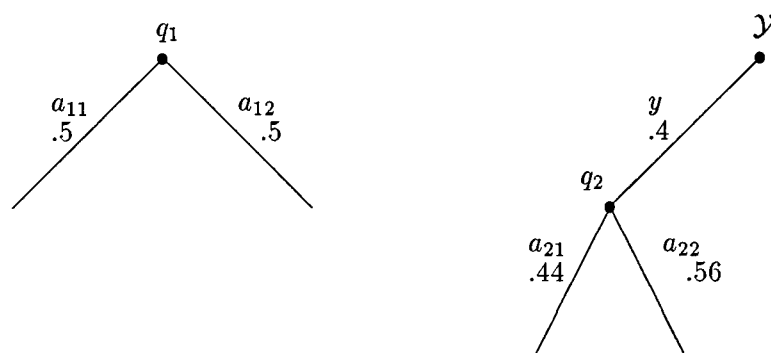


Figure 7.7: The Cluster Decomposition Terms After the Estimation Procedure.

CHAPTER 8

CONCLUSIONS AND FUTURE DIRECTIONS

This dissertation has introduced a new framework for uncertain knowledge representation in the context of stochastic decision theory. The major advantage this representation offers is the manipulation capability which is lacking in alternate representation tools such as belief networks. The manipulation rules, which are inspired from those introduced by Cockett for discrete decision theory, are sound and complete, and thus result in a consistent theory.

The terms of stochastic decision theory (stochastic terms) incorporate the probabilistic information into the representing structure. The manipulation rules then not only deal with the structure of the term but also with the probabilistic data, in conformance with the laws of probability theory. This important feature has directed the dissertation towards familiar problems in control systems theory which involve probability, such as state estimation. In this direction, a state estimation problem was re-cast from the stochastic decision theory point of view, aiming at resolving some of the obstacles encountered in large scale systems. The investigation not only introduces a new interpretation of state estimation in terms of stochastic term manipulations, but also allows, in the context of stochastic decision theory, a decomposition of a state estimation problem into a set of smaller sub-problems, alleviating difficulties in large scale system.

The cluster decomposition scheme devised for stochastic terms used very useful concepts that led to the solution of some known problems in uncertain knowledge representation such as the entailment problem and the detection of statistical independence (depen-

dence). A simple algorithm is given which solves the problem of probabilistic entailment in stochastic decision theory according to the maximum entropy principle. Another algorithm is outlined, which, for a given stochastic term, simultaneously uncovers all the independence and dependence classes. All these results are used to construct, optimally, a cluster decomposition scheme for stochastic terms.

Stochastic decision theory may be applied in a wide variety of disciplines. An obvious area of application is the representation of uncertain knowledge for knowledge-based systems. In this area, the new capabilities provided by this theory for manipulating the stochastic term structure, retrieving information from it, and updating it, provide significant improvements over the current state of the art. Real-time system applications are an attractive test for the cluster decomposition method.

The problems addressed in this dissertation suggests some areas for future research. These are

1. Stochastic decision theory with many identities: Chapter 4 is restricted to stochastic decision theories with a single identity. It is important to investigate the behavior of these theories in the presence of several identities. This can lead to a decomposition method for stochastic terms similar to the one of Chapter 7.
2. Combined random experiment modeling: Stochastic decision theory can be introduced via a purely probabilistic setting. Stochastic terms can be viewed as a random experiment which consists of a set of sequentially performed and interdependent random experiments. In this setting, it is important to investigate whether the completeness results established in Chapter 4 can be derived.
3. Conditional independence/dependence: The algorithm we described in Chapter 6 simultaneously detects all the independence and dependence classes in a given stochas-

tic term. This algorithm can theoretically be used to uncover conditional independence/dependence as well. However, the complexity grows exponentially with the number of the stochastic term decisions which makes the problem untractable. An efficient algorithm is needed for the detection of all the conditional independence/dependence relationships. A criterion for best representation can be devised, accordingly, and an optimal representation for stochastic terms can be obtained.

4. State estimation in dynamic systems and its decomposition: The state estimation problem addressed in Chapter 7 deals with finite state static systems. Dynamic systems can be similarly treated in the same context, provided that the dynamics of the system states are incorporated into the stochastic term describing the structure. This can be done by viewing these dynamics as an additional decision, the *dynamics decision*, and appending it to the system's stochastic term after each step of the estimation procedure.
5. Knowledge acquisition in expert systems: Stochastic decision theory not only provides a convenient and flexible representation tool for expert system knowledge bases, but also offers a good potential in solving open problems in the expert system area. The problem of knowledge acquisition, because of its statistical nature, may be approached from the stochastic decision theory point of view. The manipulation capability of the theory may be the key for extracting, from the examples, stochastic terms that better represent expert system knowledge.

BIBLIOGRAPHY

BIBLIOGRAPHY

- [1] Adámek, J., *Theory of Mathematical Structures*. Reidel, Prague, Czechoslovakia, 1983.
- [2] Arbib, M.A. and E.G. Manes, *Arrows, Structures, and Functors: The Categorical Imperative*. Academic Press, New York, 1975.
- [3] Athans, M. and P.L. Falb, *Optimal Control*. McGraw-Hill, New York, 1966.
- [4] Barr, M. and C. Wells, *Toposes, Triples, and Theories*. Springer-Verlag, New York, 1985.
- [5] Bloom, S. and R. Tindell, "Varieties of if ... then ... else ...," *Siam J. on Compt.*, Vol. 12, 1983.
- [6] Booth, T.L., *Sequential Machines and Automata Theory*. Wiley & Sons, New York, 1967.
- [7] Booth, T.L., "Estimation, prediction, and smoothing in discrete parameter systems," *IEEE Transactions on Computers*, Vol. C-19, No.12, December 1970, pp. 1193-1203.
- [8] Brown, D.T., "A note on approximations to discrete probability distribution," *Information and Control* 2, 1959, pp. 386-392.
- [9] Bucur, I. and A. Deleanu, *Introduction to the Theory of Categories and Functors*. Wiley & Sons, London, 1968.
- [10] Cheeseman, P. "A method of computing generalized Bayesian probability values for expert systems," *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, East Germany (William Kaufmann, ed., Los Altos, CA, 1983), pp. 198-202.
- [11] Chen, C.T., *Introduction to Linear System Theory*. Holt, Rinehart, and Winston, Inc., New York, 1970.
- [12] Chen, K. and Z. Ras "Homogenous information trees," *Fundamenta Informaticae* VIII, 1985, pp. 123-149.
- [13] Chen, S., "Evidential logic and Dempster-Shafer theory," *Proceedings of the International Symposium on Methodologies for Intelligent Systems*, Knoxville, TN, October 1986, pp. 201-206.
- [14] Chow, C.K. and C.N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Transactions on Information Theory*, Vol. IT-14, No. 3, May 1968, pp. 462-467.
- [15] Cockett, J.R.B., J. Zrida, and J.D. Birdwell, "Stochastic decision theory," To appear in *Probability in the Engineering and Information Sciences*.

- [16] Cockett, J.R.B., "Discrete decision theory: manipulations," *Theor. Comp. Sci.* 54, 1987, pp. 215-236.
- [17] Cockett, J.R.B., "Optimizing decision expressions," *Fundamenta Informaticae* X, 1987, pp. 93-114.
- [18] Cockett, J.R.B., *Discrete Decision Theory: Completeness*. Technical Report, CS-87-69, Dept. of Comp. Sc., Univ. of Tennessee at Knoxville, TN, January 1987.
- [19] Cockett, J.R.B. and J. Herrera, "Prime rules based methodologies give inadequate control," *Colloquia Proceedings of the International Symposium on Methodologies for Intelligent Systems*, Knoxville, TN, October 1986, pp. 441-449.
- [20] Cockett, J.R.B. and J. Herrera, *Decision Tree Reduction*. Technical Report, CS-85-68, Dept. of Comp. Sc., Univ. of Tennessee at Knoxville, TN, 1987.
- [21] Cohen, P.R. and E.A. Feigenbaum, *The Handbook of Artificial Intelligence*. Vol. 1, 2, and 3, William Kauffmann, Los Altos, California, 1982.
- [22] Dattatreya, G.R. and L.N. Kanal, "Adaptive pattern recognition with random costs and its application to decision trees," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-16, No. 2, March/april 1986, pp. 208-218.
- [23] Doberkat, E.E., *Stochastic Automata: Stability, Nondeterminism, and Prediction*. Springer-Verlag, New York, 1981.
- [24] Drake, A.W., *Fundamentals of Applied Probability Theory*. McGraw-Hill, New York, 1967.
- [25] Feinstein, A., *Foundations of Information Theory*. McGraw-Hill, New York, 1958.
- [26] Gelb, A., *Applied Optimal Estimation*. The M.I.T. Press, Cambridge, Massachusetts, 1974.
- [27] Goldblatt, R., *Topoi: The Categorical Analysis of Logic. Study in Logic and The Foundations of Mathematics*. Edited by J. Barwise et. al., Vol. 98, North-Holland, 1984.
- [28] Grätzer, G., *General Lattice Theory*. Academic Press, New York, 1978.
- [29] Grätzer, G., *Universal Algebra*. Springer-Verlag, New York, 1979.
- [30] Guessarian, I. and Meseguer, J., "On the axiomatization of "if-then-else"," *Siam J. on Compt.*, Vol. 16, No. 2, April 1987.
- [31] Hartmanis, J., "The application of some basic inequalities for entropy," *Information and Control* 2, 1959, pp. 199-213.

- [32] Herrera, J.A. and J.R.B. Cockett, *Finding All the Optimal Forms of a Decision Tree*. Technical Report CS-86-66, Dept. of Comp. Sci., Univ. of Tennessee at Knoxville, TN, November, 1986.
- [33] Herrera, J.A., *Theoretical Foundations and Algorithms for the Generations of Optimal Decision Trees*. Ph.D. Dissertation, Dept. of Elec. & Comp. Engr., Univ. of Tennessee at Knoxville, TN, March, 1988.
- [34] Hyafil, L., "Constructing optimal binary trees is NP-complete," *Information Processing Letters*, 5(1), 1976, pp. 15-17.
- [35] Jazwinski, A.H., *Stochastic Processes and Filtering Theory*. Academic Press, New York, 1970.
- [36] Johnstone, P.T., *Topos Theory*. Academic Press, London, 1977.
- [37] Kalman, R.E., P.L. Falb, and M.A. Arbib, *Topics in Mathematical System Theory*. McGraw-Hill, New York, 1969.
- [38] Kalman, R.E., "A new approach to linear filtering and prediction problems," *Trans. ASME, Ser. D: J. Basic Engr.* 82, 1960, pp.35-45.
- [39] Kalman, R.E. and R.S. Bucy, "New results in linear filtering and prediction theory," *Trans. ASME, Ser. D: J. Basic Engr.* 83, 1961, pp.95-108.
- [40] Kolman, B. and R.C. Busby, *Discrete Mathematical Structures for Computer Science*. Prentice-Hall, New Jersey, 1984.
- [41] Ku, H.H. and S. Kullback, "Approximating discrete probability distributions," *IEEE Transactions on Information Theory*, Vol. IT-15, No. 4, July 1969, pp. 444-447.
- [42] Kushner, H.J., "On the dynamical equations of conditional probability density functions, with applications to optimal stochastic control theory," *J. Math. Anal. Appl.* 8, 1964, pp. 332-344.
- [43] Kushner, H.J., "On the differential equations satisfied by conditional probability densities of Markov processes," *SIAM J. Control* 2, 1964, pp. 106-119.
- [44] Lewis, P.M. II, "Approximating probability distributions to reduce storage requirement," *Information and Control* 2, 1959, pp. 214-225.
- [45] Lorenz, A.A., *Stochastic Automata. Constructive Theory*. Wiley & Sons, New York, 1974.
- [46] Luenberger, D.G., *Introduction to Linear and Nonlinear Programming*. Addison-Wesley, Reading, MA, 1973.
- [47] Mac Lane, S., *Categories for Working Mathematician*. Springer-Verlag, New York, 1971.

- [48] Manes, E.G., *Algebraic Theories*. Volume 26 of GTM, Springer-Verlag, 1976.
- [49] Manes, E.G., *Category Theory Applied to Computation and Control*. Springer-Verlag, 1974.
- [50] Mitchell, B., *Theories of Categories*. Academic Press, New York, 1965.
- [51] Moret, B.M.E., M.G. Thomason, and R.C. Gonzalez, "The activity of a variable and its relation to decision trees," *ACM Transactions on Programming Languages and Systems*, Vol.2, No.4, October 1980, pp. 580-595.
- [52] Moret, B.M.E., *The Representation of Discrete Functions by Decision Trees: Aspects of Complexity and Problems of Testing*. Ph.D. Dissertation, The University of Tennessee at Knoxville, Department of Electrical Engineering, Knoxville, TN, August 1980.
- [53] Moret, B.M.E., "Decision trees and diagrams," *Computing Surveys*, Vol.14, No.4, December 1982, pp. 593-623.
- [54] Negoita, C.V., *Expert Systems and Fuzzy Systems*. Benjamin/Cumming, Menlo Park, CA, 1985.
- [55] Nilsson, N.J., "Probabilistic logic," *Artificial Intelligence* 28 (1986) 71-87.
- [56] Papoulis, A., *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, New York, 1965.
- [57] Paz, A., *Introduction to Probabilistic Automata*. Academic Press, New York, 1971.
- [58] Pearl, J., "Fusion, propagation, and structuring in belief networks," *Artificial Intelligence* 29 (1986) 241-288.
- [59] Quinlan, J.R., "Induction of decision trees," *Machine Learning*, Vol. 1, 1986, pp. 81-106.
- [60] Quinlan, J.R., *Applications of Expert Systems*. Addison-Wesley, Wokingham, England, 1987.
- [61] Revuz, D., *Markov Chains*. North-Holland, Amsterdam, The Netherlands, 1984.
- [62] Romanovsky, V.I., *Discrete Markov Chains*. Wolters-Noordhoff, The Netherlands, 1970.
- [63] Rudin, W., *Real and Complex Analysis*. McGraw-Hill, New York, 1966.
- [64] Shannon, C.E. and W. Weaver, *The Mathematical Theory of Communication*. University of Illinois Press, Chicago, IL, 1949.
- [65] Simon, J.C., *Computer Oriented Learning Processes*. Noordhoff, Leyden, The Netherlands, 1976.

- [66] Stratonovich, R.L., "On the theory of optimal nonlinear filtration of random functions," *Theor. Probability Appl.* 4, 1959, pp. 223-225.
- [67] Stratonovich, R.L., "Conditional Markov processes," *Theor. Probability Appl.* 5, 1960, pp. 156-178.
- [68] Wong, S.K.M. and W. Ziarko, "Algebraic versus probabilistic independence in decision theory," *Colloquia Proceedings of the International Symposium on Methodologies for Intelligent Systems*, Knoxville, TN, October 1986, pp. 141-147.
- [69] Zrida, J., J.D. Birdwell, and J.R.B. Cockett, "Uncertain knowledge representation via stochastic decision trees," *Proceeding of the Scientific Symposium MARI 87, Cognitiva 87/Electronic Image Electronique*, La Villette, Paris, France, May 1987, pp. 111-116.
- [70] Zrida, J., J.D. Birdwell, and J.R.B. Cockett, "State estimation for static systems defined by stochastic decision trees," *Proceedings of the 19th Southeastern Symposium on System Theory*, Clemson, South Carolina, April 1987, pp. 487-491.
- [71] Zrida, J., J.D. Birdwell, and J.R.B. Cockett, "State estimation in finite stochastic decision trees," *Colloquia Proceedings of the International Symposium on Methodologies for Intelligent Systems*, Knoxville, TN, October 1986, pp. 35-49.
- [72] Zrida, J., J.D. Birdwell, and J.R.B. Cockett, "Applying stochastic decision theory to estimation problems," *Proceedings of the 26th IEEE Conference on Decision and Control*, Los Angeles, California, December 1987, pp. 1180-1184.
- [73] Zrida, J., J.D. Birdwell, and J.R.B. Cockett, "Probabilistic entailment in stochastic decision theory," *Proceedings of the 1988 Technical Symposium Southeast on Optics, Electro-Optics, and Sensors*, Orlando, Florida, April 1988.
- [74] Zrida, J. and J.D. Birdwell, "Statistical independence in stochastic decision theory," In Preparation.

APPENDICES

APPENDIX A

APPENDIX TO CHAPTER 3

A.1 APPENDIX TO SECTION 3.3

A.1.1 Proof of Theorem 4 on Page 59

The proof consists of the following cyclic implications $(i) \Rightarrow (ii) \Rightarrow (iii) \Rightarrow (i)$.

($(i) \Rightarrow (ii)$)

Obvious by definition.

($(ii) \Rightarrow (iii)$)

Obvious.

($(iii) \Rightarrow (i)$)

We need to show that if, for a given ordering $l_1, l_2, \dots, l_n$ of these n events, we have

$$\begin{aligned} p(A_{l_1}, A_{l_2}) &= p(A_{l_1})p(A_{l_2}) \\ p(A_{l_1}, A_{l_2}, A_{l_3}) &= p(A_{l_1})p(A_{l_2})p(A_{l_3}) \\ &\vdots \\ p(A_{l_1}, A_{l_2}, A_{l_3}, \dots, A_{l_n}) &= p(A_{l_1})p(A_{l_2})p(A_{l_3}) \cdots p(A_{l_n}), \end{aligned}$$

then, for all distinct indices $i, j, \dots, k \leq n$, we have that

$$p(A_i, A_j, \dots, A_k) = p(A_i)p(A_j) \cdots p(A_k).$$

Clearly,

$$\begin{aligned}
 p(A_i, A_j, \dots, A_k) &= \underbrace{\sum \cdots \sum}_{\text{non-matching indices}} p(A_{l_1}, A_{l_2}, \dots, A_{l_n}) \\
 &= \underbrace{\sum \cdots \sum}_{\text{non-matching indices}} p(A_{l_1})p(A_{l_2}) \cdots p(A_{l_n}) \\
 &= p(A_i)p(A_j) \cdots p(A_k).
 \end{aligned}$$

APPENDIX B

APPENDIX TO CHAPTER 4

B.1 APPENDIX TO SECTION 4.2

B.1.1 Proof of Lemma 6 on Page 70

If the pullback exists, then it is a direct consequence of the factorization that the pullback must be a measure-preserving monic. It suffices thus to show that the pullback exists. A measure-preserving monic is determined by the image of the $\top$ of the sub-description space. Using the boolean map of the description map over which the pullback is being formed, allows this $\top$ to be carried into the domain of the description map. This gives a hereditary sub-object of the domain of this description map. It is easy to check that this indeed gives a pullback of description spaces.

B.1.2 Proof of Lemma 7 on Page 71

We shall show that $\mathcal{C}$ has finite coproducts. Let $X_1 = (\mathcal{B}_1, m_1)$ and $X_2 = (\mathcal{B}_2, m_2)$ be objects of the category $\mathcal{C}$, then define $X_1 \oplus X_2$ to be $(\mathcal{B}_1 \otimes \mathcal{B}_2, m_1 \oplus m_2)$, where

$$(m_1 \oplus m_2)(B_1, B_2) = m_1(B_1) + m_2(B_2).$$

It is easy to see that

$$\begin{aligned} b_1 : X_1 &\longrightarrow X_1 \oplus X_2 \\ B_1 &\longmapsto (B_1, \perp), \text{ and} \\ b_2 : X_2 &\longrightarrow X_1 \oplus X_2 \\ B_2 &\longmapsto (\perp, B_2) \end{aligned}$$

are measure-preserving monics on X_1 and X_2 , respectively. The hereditary subobjects are generated by, respectively, $(\top, \perp)$ and $(\perp, \top)$.

Suppose that, for an arbitrary object $X_0 = (\mathcal{B}_0, m_0)$, $f : X_1 \longrightarrow X_0$ and $g : X_2 \longrightarrow X_0$ are description space maps, then we have the boolean algebra maps $f^* : \mathcal{B}_0 \longrightarrow \mathcal{B}_1$ and $g^* : \mathcal{B}_0 \longrightarrow \mathcal{B}_2$ which give a unique map to the product boolean algebras, namely $\langle f^*, g^* \rangle : \mathcal{B}_0 \longrightarrow \mathcal{B}_1 \otimes \mathcal{B}_2$. The map we require is the description map which corresponds to this, which is simply the join of the original maps in $\mathcal{B}_0$. Thus, the fact that it is a coproduct follows from the fact that it is a product in boolean algebras, as does the fact that they are disjoint. Furthermore, the map does not increase measure as the measure of a joint image is bounded by the sum of the measures of the individual images.

B.2 APPENDIX TO SECTION 4.3

B.2.1 Proof of Lemma 8 on Page 72

Given a decision

$$q : S \longrightarrow S \oplus S \oplus \cdots \oplus S,$$

if $r : R \longrightarrow S$ is a measure-preserving monic on the subobject R of S , then the right-hand square of the following diagram is a pullback of the map r and codiagonal map ∇_S of the object S .

$$\begin{array}{ccccc}
& & \text{Tr}(q) = i(S) & & \\
& \swarrow & & \searrow & \\
S & \xrightarrow{q} & S \oplus \cdots \oplus S & \xrightarrow{\nabla_S} & S \\
\uparrow r & & \uparrow r \oplus \cdots \oplus r & & \uparrow r \\
R & \xrightarrow{q \mid r} & R \oplus \cdots \oplus R & \xrightarrow{\nabla_R} & R \\
& \swarrow & & \searrow & \\
& & \text{Tr}(q \mid r) = i(R) & &
\end{array}$$

The object R , with maps $r.q$ and $i(R)$, is a candidate pullback for the right square and thus has a comparison map to $R \oplus \cdots \oplus R$. This map has been denoted $q \mid r$. However, the outer square is clearly a pullback. Thus, the leftmost square must be a pullback. Furthermore, $q \mid r$ is a decision as $(q \mid r).\nabla_R$ is the identity.

B.2.2 Proof of Theorem 6 on Page 77

Using [SD.2] for each variable with zero probability and substituting, at each occurrence the “ $-$ ” by x_i , we obtain

$$\begin{aligned}
q : [a_1/0 \gg x_1, \dots, a_i/1 \gg x_i, \dots, a_m/0 \gg x_m] \\
& \iff \\
q : [a_1/0 \gg x_i, \dots, a_i/1 \gg x_i, \dots, a_m/0 \gg x_i] \\
& \implies \\
& x_i,
\end{aligned}$$

where the last step uses [SD.1].

B.2.3 Proof of Theorem 7 on Page 78

1) The proof is by induction on the height of these composite terms. The **height** of a composite term is defined as the number of decisions in the longest path in the term. If the height of w is 1, then w is a primitive decision and the result holds by [SD.1]. Suppose that the result hold for any stochastic term of height less than or equal to $n - 1$. If w is a stochastic term of height n , then there exists a decision $q \in Q$ such that

$$\begin{aligned} w : [x_1/p_1 \gg x, \dots, x_m/p_m \gg x] \\ = \\ q : [a_1/r_1 \gg w_1 : [x_{11}/p_{11} \gg x, \dots, x_{1m}/p_{1m} \gg x], \dots \\ \dots, a_m/r_m \gg w_m : [x_{m1}/p_{m1} \gg x, \dots, x_{mm}/p_{mm} \gg x]], \end{aligned}$$

where the w_i 's are stochastic terms of height at most $n - 1$. For these stochastic terms, idempotence is assumed to hold, and when applied to each w_i yields

$$w : [x_1/p_1 \gg x, \dots, x_m/p_m \gg x] \implies q : [a_1/r_1 \gg x, \dots, a_m/r_m \gg x] \implies x,$$

where the last implication is due to [SD.1], as q is a primitive decision. Thus, idempotence holds for stochastic terms as well.

2) Again by induction on the height of the stochastic term w . If the height of w is 1, then w is a primitive decision and the result holds by [SD.2]. Suppose that the result hold for any stochastic term of height less than or equal to $n - 1$. If w is a stochastic term of height n , then there exists a decision $q \in Q$ such that

$$\begin{aligned} w : [x_1/p_1 \gg t_1, \dots, x_i/0 \gg t_i, \dots, x_m/p_m \gg t_m] \\ = \\ q : [a_1/r_1 \gg w_1 : [x_{11}/p_{11} \gg t_{11}, \dots, x_{1i}/p_{1i} \gg t_{1i}, \dots, x_m/p_{1m} \gg t_{1m}], \dots \\ \dots, a_m/r_m \gg w_m : [x_1/p_{m1} \gg t_{m1}, \dots, x_i/p_{mi} \gg t_{mi}, \dots, x_m/p_{mm} \gg t_{mm}]], \end{aligned}$$

where the w_i 's are stochastic terms of height at most $n - 1$ exactly one of which (say w_{i_0}) must have a subtree $t_{i_0j} = t_i$ and $p_{i_0j} = 0$. Use the hypothesized indifference to replace, in w_{i_0} , t_{i_0j} (which is in fact t_i) by “ $-$ ”. Reversing the above operations leads to

$$w : [x_1/p_1 \gg t_1, \dots, x_i/0 \gg -, \dots, x_m/p_m \gg t_m],$$

which is the desired result.

3) Again by induction on the height of the stochastic term w . If the height of w is 1, then w is a primitive decision and the result holds by [SD.3]. Suppose that the result hold for any stochastic term of height less than or equal to $n - 1$. If w is a stochastic term of height n , then there exists a decision $q \in Q$ such that

$$\begin{aligned} w : [x_1/p_1 \gg t_1, \dots, x_i/p_i \gg t_i, \dots, x_m/p_m \gg t_m] \\ = \\ q : [a_1/r_1 \gg w_1 : [x_{11}/p_{11} \gg t_1, \dots, x_{1i}/p_{1i} \gg t_i, \dots, x_{1m}/p_{1m} \gg t_m], \dots \\ \dots, a_i/r_i \gg w_i : [x_{i1}/p_{i1} \gg t_1, \dots, x_{ii}/p_{ii} \gg t_i, \dots, x_{im}/p_{im} \gg t_m], \dots \\ \dots, a_m/r_m \gg w_m : [x_{m1}/p_{m1} \gg t_1, \dots, x_{mi}/p_{mi} \gg t_i, \dots, x_{mm}/p_{mm} \gg t_m]], \end{aligned}$$

where the w_i 's are stochastic terms of height at most $n - 1$. Use the repetition rule [SD.3] for the primitive decision q to repeat that decision at every one of its leaves to obtain that the previous expression is equivalent to

$$\begin{aligned} q : [a_1/r_1 \gg q : [a_1/1 \gg w_1 : [x_{11}/p_{11} \gg t_1, \dots, x_{1i}/p_{1i} \gg t_i, \dots \\ \dots, x_{1m}/p_{1m} \gg t_m], \dots, a_i/0 \gg -, \dots, a_m/0 \gg -], \dots \\ \dots, a_i/r_i \gg q : [a_i/0 \gg -, \dots, a_i/1 \gg w_i : [x_{i1}/p_{i1} \gg t_1, \dots \\ \dots, x_{ii}/p_{ii} \gg t_i, \dots, x_{im}/p_{im} \gg t_m], \dots, a_m/0 \gg -], \dots \\ \dots, a_m/r_m \gg q : [a_m/0 \gg -, \dots, a_m/1 \gg w_m : [x_{m1}/p_{m1} \gg t_1, \dots \\ \dots, x_{mi}/p_{mi} \gg t_i, \dots, x_{mm}/p_{mm} \gg t_m]]]. \end{aligned}$$

We then use the transposition manipulation [SD.4], for composite terms, provided that it holds, which will be shown next, to put each of the composite term w_i 's above the decision q and then use the certainty manipulation [SD.5], for terms, to remove the decision q except at the i^{th} leaf of every w_i , and obtain that the previous expression is equivalent (along the left-to-right way only) to

$$\begin{aligned}
& q : [a_1/r_1 \gg w_1 : [x_{11}/p_{11} \gg t_1, \dots, x_{1i}/p_{1i} \gg q : [a_1/1 \gg t_i, \dots, a_i/0 \gg - , \dots \\
& \dots, a_m/0 \gg -], \dots, x_{1m}/p_{1m} \gg t_m], \dots, a_i/r_i \gg w_i : [x_{i1}/p_{i1} \gg t_1, \dots \\
& \dots, x_{ii}/p_{ii} \gg q : [a_1/0 \gg - , \dots, a_i/1 \gg t_i, \dots, a_m/0 \gg -], \dots \\
& \dots, x_{im}/p_{im} \gg t_m,], \dots, a_m/r_m \gg w_m : [x_{m1}/p_{m1} \gg t_1, \dots \\
& \dots, x_{mi}/p_{mi} \gg q : [a_1/0 \gg - , \dots, a_i/0 \gg - , \dots \\
& \dots, a_m/1 \gg t_i], \dots, x_{mm}/p_{mm} \gg t_m]].
\end{aligned}$$

We now use the hypothetized repetition to repeat each one of the w_i 's at its i^{th} leaf to obtain that the previous expression is in turn equivalent to

$$\begin{aligned}
& q : [a_1/r_1 \gg w_1 : [x_{11}/p_{11} \gg t_1, \dots, x_{1i}/p_{1i} \gg w_1 : [x_{11}/0 \gg - , \dots \\
& \dots, x_{1i}/1 \gg q : [a_1/1 \gg t_i, \dots, a_i/0 \gg - , \dots, a_m/0 \gg -], \dots, x_{1m}/0 \gg -], \dots \\
& \dots, x_{1m}/p_{1m} \gg t_m], \dots, a_i/r_i \gg w_i : [x_{i1}/p_{i1} \gg t_1, \dots \\
& \dots, x_{ii}/p_{ii} \gg w_i : [x_{i1}/0 \gg - , \dots, x_{ii}/1 \gg q : [a_1/0 \gg - , \dots, a_i/1 \gg t_i, \dots \\
& \dots, a_m/0 \gg -], \dots, x_{im}/0 \gg -], \dots, x_{im}/p_{im} \gg t_m,], \dots \\
& \dots, a_m/r_m \gg w_m : [x_{m1}/p_{m1} \gg t_1, \dots, x_{mi}/p_{mi} \gg w_m : [x_{m1}/0 \gg - , \dots \\
& \dots, x_{mi}/1 \gg q : [a_1/0 \gg - , \dots, a_i/0 \gg - , \dots, a_m/1 \gg t_i], \dots \\
& \dots, x_{mm}/0 \gg -], \dots, x_{mm}/p_{mm} \gg t_m]].
\end{aligned}$$

Composing back q and the w_i 's, the previous expression is identified to

$$\begin{aligned}
& w : [x_1/p_1 \gg t_1, \dots, x_i/p_i \gg w : [x_1/0 \gg t_1, \dots, x_i/1 \gg t_i, \dots \\
& \dots, x_m/0 \gg t_m], \dots, x_m/p_m \gg t_m]
\end{aligned}$$

This proves the left-to-right implication of [SD.3]. The right-to-left implication is obvious if we use [SD.2] in which we let $u_i = t_i$ and then apply [SD.1].

4) The proof is a double induction on the height of two stochastic term w_1 and w_2 . If both of these terms have a height of 1, the transposition manipulation [SD.4] provides the result. We first need to show that any stochastic term, w_1 , of an arbitrary height, n , can be transposed with any stochastic term w_2 of height 1 (primitive term). This proof is again by induction on the height of the stochastic terms w_1 . Suppose that the result hold for any stochastic term of height less than or equal to $n - 1$. Suppose now that w_1 has height n . Then, given w_1 , there exists a decision $q \in Q$ such that

$$\begin{aligned} & w_1 : [x_1/p_1 \gg w_2 : [x_{11}/p_{11} \gg t_{11}, \dots, x_{1m_2}/p_{1m_2} \gg t_{1m_2}], \dots \\ & \dots, x_{m_1}/p_{m_1} \gg w_2 : [x_{11}/p_{m_11} \gg t_{m_11}, \dots, x_{1m_2}/p_{m_1m_2} \gg t_{m_1m_2}]] \\ & = \\ & q : [a_1/p'_1 \gg w_{11} : [x_1/s_{11} \gg w_2 : [x_{11}/p_{11} \gg t_{11}, \dots, x_{1m_2}/p_{1m_2}], \dots \\ & \dots, x_{m_1}/s_{1m_1} \gg w_2 : [x_{11}/p_{m_11} \gg t_{m_11}, \dots, x_{1m_2}/p_{m_1m_2} \gg t_{m_1m_2}]], \dots \\ & \dots, a_{m_1}/p'_{m_1} \gg w_{1m_1} : [x_1/s_{m_11} \gg w_2 : [x_{11}/p_{11} \gg t_{11}, \dots, x_{1m_2}/p_{1m_2} \gg t_{1m_2}], \dots \\ & \dots, x_{m_1}/s_{m_1m_1} \gg w_2 : [x_{11}/p_{m_11} \gg t_{m_11}, \dots, x_{1m_2}/p_{m_1m_2} \gg t_{m_1m_2}]]], \end{aligned}$$

where the w_{1j} 's have a height of at most $n - 1$ and

$$\sum_{j=1}^{m_1} p'_j s_{jk} = p_k \quad \text{for } k = 1, \dots, m_1.$$

Now, for each $j = 1, \dots, m_1$, transposing w_{1j} and w_2 , the previous expression is equivalent to

$$\begin{aligned} & q : [a_1/p'_1 \gg w_2 : [x_{11}/u_{11} \gg w_{11} : [x_1/v_{11}^1 \gg t_{11}, \dots, x_{m_1}/v_{1m_1}^1 \gg t_{m_11}], \dots \\ & \dots, x_{1m_2}/u_{1m_2} \gg w_{11} : [x_1/v_{m_21}^1 \gg t_{1m_2}, \dots, x_{m_1}/v_{m_2m_1}^1 \gg t_{m_1m_2}]], \dots \\ & \dots, a_{m_1}/p'_{m_1} \gg w_2 : [x_{11}/u_{m_11} \gg w_{1m_1} : [x_1/v_{11}^{m_1} \gg t_{11}, \dots, x_{m_1}/v_{1m_1}^{m_1} \gg t_{m_11}], \dots \\ & \dots, x_{1m_2}/u_{m_1m_2} \gg w_{1m_1} : [x_1/v_{m_21}^{m_1} \gg t_{1m_2}, \dots, x_{m_1}/v_{m_2m_1}^{m_1} \gg t_{m_1m_2}]]], \end{aligned}$$

with

$$\begin{cases} u_{ki} &= \sum_{j=1}^{m_1} s_{kj} p_{ji} \\ v_{ij}^k &= \frac{s_{kj} p_{ji}}{u_{ki}}, \end{cases}$$

where $i = 1, \dots, m_2$, and $j, k = 1, \dots, m_1$.

Using [SD.4] to transpose q and w_2 , the previous expression is in turn equivalent to

$$\begin{aligned} w_2 : [x_{11}/r_1 \gg q : [a_1/l_{11} \gg w_{11} : [x_1/v_{11}^1 \gg t_{11}, \dots, x_{m_1}/v_{1m_1}^1 \gg t_{m_11}], \dots \\ \dots, a_{m_1}/l_{1m_1} \gg w_{1m_1} : [x_1/v_{11}^{m_1} \gg t_{11}, \dots, x_{m_1}/v_{1m_1}^{m_1} \gg t_{m_11}]]], \dots \\ \dots, x_{1m_2}/r_{m_2} \gg q : [a_1/l_{m_21} \gg w_{11} : [x_1/v_{m_21}^1 \gg t_{1m_2}, \dots, x_{m_1}/v_{m_2m_1}^1 \gg t_{m_1m_2}], \dots \\ \dots, a_{m_1}/l_{m_2m_1} \gg w_{1m_1} : [x_1/v_{m_21}^{m_1} \gg t_{1m_2}, \dots, x_{m_1}/v_{m_2m_1}^{m_1} \gg t_{m_1m_2}]]], \end{aligned}$$

with

$$\begin{cases} r_i &= \sum_{j=1}^{m_1} p'_j u_{ji} \\ l_{ij} &= \frac{p'_j u_{ji}}{r_i}, \end{cases}$$

where $i = 1, \dots, m_2$, and $j = 1, \dots, m_1$.

Grouping q and the w_{1j} 's back together, the previous expression is in fact no other than

$$\begin{aligned} w_2 : [x_{11}/r_1 \gg w_1 : [x_1/r_{11} \gg t_{11}, \dots, x_{m_1}/r_{1m_1} \gg t_{m_11}], \dots \\ \dots, w_1 : [x_1/r_{m_21} \gg t_{1m_2}, \dots, x_{m_1}/r_{m_2m_1} \gg t_{m_1m_2}]], \end{aligned}$$

with

$$r_{ij} = \sum_{k=1}^{m_1} l_{ik} v_{ij}^k = \sum_{k=1}^{m_1} \left(\frac{p'_k u_{ki}}{r_i} \times \frac{s_{kj} p_{ji}}{u_{ki}} \right) = \left(\frac{p_{ji}}{r_i} \right) \times \sum_{k=1}^{m_1} p'_k s_{kj} = \frac{p_{ji} p_j}{r_i},$$

and

$$r_i = \sum_{j=1}^{m_1} p'_j u_{ji} = \sum_{j=1}^{m_1} \left\{ p'_j \times \left[\sum_{k=1}^{m_1} s_{jk} p_{ki} \right] \right\} = \sum_{j=1}^{m_1} \sum_{k=1}^{m_1} p'_j s_{jk} p_{ki} = \sum_{k=1}^{m_1} p_k p_{ki},$$

where $i = 1, \dots, m_2$, and $j = 1, \dots, m_1$. This is the desired result. The induction must be completed by allowing the composite term w_2 to have an arbitrary height. This can be easily done following the same procedures.

B.2.4 Proof of Corollary 1 on Page 79

($\Leftarrow$)

Using the certainty manipulation [SD.5] for the sub-term rooted at w , we obtain that

$$\begin{aligned}
 & q : [a_1/p_1 \gg t_1, \dots, a_i/p_i \gg w : [x_1/r_1 \gg X_1, \dots \\
 & \dots, x_j/r_j \gg q : [a_1/0 \gg t_1, \dots, a_i/1 \gg X_j, \dots, a_m/0 \gg t_m], \dots \\
 & \dots, x_n/r_n \gg X_n], \dots, a_m/p_m \gg t_m] \\
 & \Rightarrow \\
 & q : [a_1/p_1 \gg t_1, \dots, a_i/p_i \gg w : [x_1/r_1 \gg X_1, \dots, x_j/r_j \gg X_j, \dots \\
 & \dots, x_n/r_n \gg X_n], \dots, a_m/p_m \gg t_m]
 \end{aligned}$$

($\Rightarrow$)

Using the repetition manipulation [SD.3], the decision q is repeated at its i^{th} argument to obtain that

$$\begin{aligned}
 & q : [a_1/p_1 \gg t_1, \dots, a_i/p_i \gg w : [x_1/r_1 \gg X_1, \dots, x_j/r_j \gg X_j, \dots \\
 & \dots, x_n/r_n \gg X_n], \dots, a_m/p_m \gg t_m] \\
 & \Leftrightarrow \\
 & q : [a_1/p_1 \gg t_1, \dots, a_i/p_i \gg q : [a_1/0 \gg t_1, \dots, a_i/1 \gg w : [x_1/r_1 \gg X_1, \dots \\
 & \dots, x_j/r_j \gg X_j, \dots, x_n/r_n \gg X_n], \dots, a_m/0 \gg t_m], \dots, a_m/p_m \gg t_m]
 \end{aligned}$$

Using the indifference manipulation [SD.3], for composite terms, the composite decision

w is duplicated at the remaining leaves of the repeated decision q . We thus obtain that

$$\begin{aligned}
& q : [a_1/p_1 \gg t_1, \dots, a_i/p_i \gg q : [a_1/0 \gg t_1, \dots, a_i/1 \gg w : [x_1/r_1 \gg X_1, \dots \\
& \quad \dots, x_j/r_j \gg X_j, \dots, x_n/r_n \gg X_n], \dots, a_m/0 \gg t_m], \dots, a_m/p_m \gg t_m] \\
& \quad \Longleftrightarrow \\
& q : [a_1/p_1 \gg t_1, \dots, a_i/p_i \gg q : [a_1/0 \gg w : [x_1/r_1 \gg X_1, \dots, x_j/r_j \gg X_j, \dots \\
& \quad \dots, x_n/r_n \gg X_n], \dots, a_i/1 \gg w : [x_1/r_1 \gg X_1, \dots, x_j/r_j \gg X_j, \dots \\
& \quad \dots, x_n/r_n \gg X_n], a_m/0 \gg w : [x_1/r_1 \gg X_1, \dots, x_j/r_j \gg X_j, \dots \\
& \quad \dots, x_n/r_n \gg X_n]], \dots, a_m/p_m \gg t_m]
\end{aligned}$$

We now use the transposition manipulation [SD.4], for composite terms, to move the composite term w right above the repeated decision q to obtain that

$$\begin{aligned}
& q : [a_1/p_1 \gg t_1, \dots, a_i/p_i \gg q : [a_1/0 \gg w : [x_1/r_1 \gg X_1, \dots, x_j/r_j \gg X_j, \dots \\
& \quad \dots, x_n/r_n \gg X_n], \dots, a_i/1 \gg w : [x_1/r_1 \gg X_1, \dots, x_j/r_j \gg X_j, \dots \\
& \quad \dots, x_n/r_n \gg X_n], \dots, a_m/0 \gg w : [x_1/r_1 \gg X_1, \dots, x_j/r_j \gg X_j, \dots \\
& \quad \dots, x_n/r_n \gg X_n]], \dots, a_m/p_m \gg t_m] \\
& \quad \Longleftrightarrow \\
& q : [a_1/p_1 \gg t_1, \dots, a_i/p_i \gg w : [x_1/r_1 \gg q : [a_1/0 \gg X_1, \dots, a_i/1 \gg X_1, \dots \\
& \quad \dots, a_m/0 \gg X_1], \dots, x_j/r_j \gg q : [a_1/0 \gg X_j, \dots, a_i/1 \gg X_j, \dots \\
& \quad \dots, a_m/0 \gg X_j], \dots, x_n/r_n \gg q : [a_1/0 \gg X_n, \dots, a_i/1 \gg X_n, \dots \\
& \quad \dots, a_m/0 \gg X_n]], \dots, a_m/p_m \gg t_m] \\
& \quad \Rightarrow \\
& q : [a_1/p_1 \gg t_1, \dots, a_i/p_i \gg w : [x_1/r_1 \gg X_1, \dots \\
& \quad \dots, x_j/r_j \gg q : [a_1/0 \gg t_1, \dots, a_i/1 \gg X_j, \dots, a_m/0 \gg t_m], \dots \\
& \quad \dots, x_n/r_n \gg X_n], \dots, a_m/p_m \gg t_m],
\end{aligned}$$

where the last transformation is due to either idempotence ([SD.1]) or certainty ([SD.5]) for composite terms.

B.3 APPENDIX TO SECTION 4.4

B.3.1 Proof of Theorem 8 on Page 82

1) If t' is obtained from t by [SD.1], then t can be written as

$$t = q : [a_1/p_1 \gg t', \dots, a_n/p_n \gg t'] .$$

If t' is a variable, then by definition X is consistent with t' . If t' is not a variable, the fact that X is consistent with t implies that X_i is consistent with t' , for $i = 1, \dots, n$. As X is isomorphic to the coproduct (disjoint union) $X_1 \oplus X_2 \oplus \dots \oplus X_n$, and each X_i is consistent with t' , X is consistent with t' .

2) If t' is the equivalent of t by [SD.2], then t can be written as

$$t = q : [a_1/p_1 \gg t_1, \dots, a_i/0 \gg t_i, \dots, a_n/p_n \gg t_n]$$

and t' as

$$t' = q : [a_1/p_1 \gg t_1, \dots, a_i/0 \gg -, \dots, a_n/p_n \gg t_n] .$$

Suppose that X satisfies the stochastic term t , then replacing t' by a variable will certainly result in a term which is satisfied by X . The difficulty lies in showing the converse: that is if t' is satisfied by X then so will t . For X to satisfy the term t , it suffices to show that any stochastic term is satisfied by X_i . Since $m_X(\top_X) \times 0 = 0 = m_{X_i}(\top_{X_i})$, every identity in the theory is satisfied by X_i as the latter has zero measure. As a result, X satisfies t .

3) If t' is equivalent to t by [SD.3], then t can be written as

$$t = q : [a_1/p_1 \gg t_1, \dots, a_i/p_i \gg t_i, \dots, a_n/p_n \gg t_n]$$

and t' as

$$t' = q : [a_1/p_1 \gg t_1, \dots, a_i/p_i \gg q : [a_1/0 \gg t_1, \dots, a_i/1 \gg t_i, \dots \\ \dots, a_n/0 \gg t_n], \dots, a_n/p_n \gg t_n]$$

Let t_0 be

$$t_0 = q : [a_1/0 \gg t_1, \dots, a_i/1 \gg t_i, \dots, a_n/0 \gg t_n].$$

Suppose that X satisfies t , for X to satisfy t' , the X_j 's, for $j \neq i$, have to be consistent with the t_j 's, which they are as X satisfies t . Furthermore, X_i must satisfy t_0 . This happens when $m_{X_i}(\top_{X_i}) \times 1 = m_{(X_i)_i}(\top_{(X_i)_i})$, which is true as $(X_i)_i$ is X_i , which follows from the fact that $q \mid \pi_i = b_i$. Therefore, X satisfies t' .

Conversely, suppose that X satisfies t' . To show that X satisfies t , we only need to show that X_i satisfies t_i . Since X_i satisfies t_0 , we have that $(X_i)_i$ satisfies t_i and $m_{X_i}(\top_{X_i}) \times 1 = m_{(X_i)_i}(\top_{(X_i)_i})$ and as a result $(X_i)_i$ is almost all of X_i . Therefore, X_i satisfies t_i , which proves that X satisfies t .

4) If t and t' are equivalent by [SD.4], then t can be written as

$$t = q_1 : [a_{11}/p_1 \gg q_2 : [a_{21}/p_{11} \gg x_{11}, \dots, a_{2m_2}/p_{1m_2} \gg x_{1m_2}], \dots \\ \dots, a_{1m_1}/p_{m_1} \gg q_2 : [a_{21}/p_{m_11} \gg x_{m_11}, \dots, a_{2m_2}/p_{m_1m_2} \gg x_{m_1m_2}]],$$

and t' as

$$t' = q_2 : [a_{21}/r_1 \gg q_1 : [a_{11}/r_{11} \gg x_{11}, \dots, a_{1m_1}/r_{1m_1} \gg x_{1m_1}], \dots \\ \dots, a_{2m_2}/r_{m_2} \gg q_1 : [a_{11}/r_{m_21} \gg x_{1m_2}, \dots, a_{1m_1}/r_{m_2m_1} \gg x_{m_1m_2}]],$$

with

$$\left\{ \begin{array}{l} r_i = \sum_{j=1}^{m_1} p_j p_{ji} \\ r_{ik} = \frac{p_k p_{ki}}{\sum_{j=1}^{m_1} p_j p_{ji}} = \frac{p_k p_{ki}}{r_i}, \end{array} \right.$$

where $i = 1, \dots, m_2$, and $j, k = 1, \dots, m_1$.

Suppose that X satisfies t . We need to show that it satisfies t' as well. Since X satisfies t , we have that

$$m_X(\top_X)p_j = m_{X_j}(\top_{X_j}), \quad \forall j = 1, \dots, m_1,$$

and also

$$m_{X_j}(\top_{X_j})p_{ji} = m_{(X_j)_i}(\top_{(X_j)_i}), \quad \forall j = 1, \dots, m_1, \text{ and } i = 1, \dots, m_2.$$

We then have the following equalities:

$$m_X(\top_X)p_jp_{ji} = m_{(X_j)_i}(\top_{(X_j)_i}), \quad \forall j = 1, \dots, m_1, \text{ and } i = 1, \dots, m_2.$$

These equalities give the consistency constraints for the paths of the stochastic term t as obtained from the usual consistency constraints written for the branches. As the transposition manipulation, [SD.4], does not affect paths and only alters the decisions on these paths, these consistency constraints do hold for the transposition-equivalent term t' of t . To obtain the consistency constraints for the branches in t' , let's form the following equalities by summing over the index j the above equalities for $i = 1, \dots, m_2$:

$$m_X(\top_X)(\sum_j^{m_1} p_j p_{ji}) = \sum_j^{m_1} m_{(X_j)_i}(\top_{(X_j)_i}) = m_{Y_i}(\top_{Y_i}), \quad \forall i = 1, \dots, m_2.$$

Since $r_i = \sum_j^{m_1} p_j p_{ji}$, these equalities become

$$m_X(\top_X)r_i = \sum_j^{m_1} m_{(X_j)_i}(\top_{(X_j)_i}) = m_{Y_i}(\top_{Y_i}), \quad \forall i = 1, \dots, m_2.$$

Also, by the previous equality,

$$m_{Y_i}(\top_{Y_i}) = m_X(\top_X)r_i = \frac{m_{(X_j)_i}(\top_{(X_j)_i})}{p_j p_{ji}} r_i$$

This means that

$$m_{Y_i}(\top_{Y_i}) \times \frac{p_j p_{ji}}{r_i} = m_{(X_j)_i}(\top_{(X_j)_i}), \quad \forall j = 1, \dots, m_1, \text{ and } i = 1, \dots, m_2.$$

Since $r_{ij} = \frac{p_j p_{ji}}{r_i}$,

$$m_{Y_i}(\top_{Y_i}) r_{ij} = m_{(X_j)_i}(\top_{(X_j)_i}), \quad \forall j = 1, \dots, m_1, \text{ and } i = 1, \dots, m_2.$$

We thus checked that all the consistency constraints for the stochastic term t' are satisfied.

Therefore X satisfies the stochastic term t' as well.

B.3.2 Proof of Theorem 9 on Page 83

($\Leftarrow$)

If X is consistent with the stochastic term t , then t is either a variable and thus is naturally a repeat-consistent term, or t can be written as

$$q : [a_1/p_1 \gg t_1, \dots, a_i/p_i \gg t_i, \dots, a_n/p_n \gg t_n]$$

where the t_i 's are stochastic terms satisfied by the objects X_i . Suppose that one of the t_i 's (say t_{i_0}) repeat the decision q , at some level. Sitting on a particular occurrence of a repeated decision, either there exists a zero probability on the path, in which case everything is consistent, or $((\dots(X_{i_1})_{i_2} \dots)_{i_{r-1}})_{i_r}$ has a non-zero measure and

$$((\dots(X_{i_1})_{i_2} \dots)_{i_{r-1}})_{i_r} = (((\dots(X_{i_1})_{i_2} \dots)_{i_{r-1}})_{i_r})_{i_{r+1}}.$$

This implies that, in the consistency constraint of the i^{th} argument

$$\begin{aligned} & m_{((\dots(X_{i_1})_{i_2} \dots)_{i_{r-1}})_{i_r}}(\top_{((\dots(X_{i_1})_{i_2} \dots)_{i_{r-1}})_{i_r}}) p_i \\ &= \\ & m_{(((\dots(X_{i_1})_{i_2} \dots)_{i_{r-1}})_{i_r})_{i_{r+1}}}(\top_{(((\dots(X_{i_1})_{i_2} \dots)_{i_{r-1}})_{i_r})_{i_{r+1}}}) \end{aligned}$$

the probability p_i must be 1. Therefore, the remaining probabilities must be all zeros. Thus, for the repeated decision, the probabilities are zeros everywhere except on the appropriate argument, where the probability is 1. This proves that t is a repeat-consistent stochastic term.

($\Rightarrow$)

Let t be a stochastic repeat-consistent term and $\mathcal{P}$ the set of all non-zero probability paths in t . Let $\mathcal{B}$ be the boolean algebra on the set of subsets of $\mathcal{P}$ and let m be the positive measure defined as follows:

$$\forall P_1, P_2, \dots, P_m \in \mathcal{P}, \quad m(\{P_1, P_2, \dots, P_m\}) = \sum_{i=1}^m p(P_i).$$

where $p(P_i)$ is the probability of the path P_i . Denote the obtained description space as $X = (\mathcal{B}, m)$. Each decision q_i in the stochastic term t is interpreted as a decision from X to the m_i -fold coproduct of X , where m_i is the arity of q_i . The interpretation of q_i places a path, in which q_i occurs, in the argument corresponding to the branch associated with q_i in that path. If q_i does not occur the path is put in the first argument. This recipe provides a method of splitting any set $B = \{P_1, P_2, \dots, P_k\}$ of the boolean algebra $\mathcal{B}$ into m_i disjoint subsets. The fact that the stochastic term t is repeat-consistent guarantees that, along a given path, only a single branch of a given decision will be selected. The obtained decision coalgebra, X , together with the interpretation of the decisions is, by definition, a model over a non-trivial full description space which satisfies t .

B.4 APPENDIX TO SECTION 4.5

B.4.1 Proof of Lemma 9 on Page 85

If the decision q is semi-essential in t , the proof of Cockett [16,17], developed for the deterministic case, applies provided that the stochastic version of the transposition manipulation, [SD.4] is used. If the decision q is reachable semi-essential in t , use the indifference rule [SD.2] to put q where it is missing to make q semi-essential.

B.4.2 Proof of Theorem 10 on Page 87

($\Leftarrow$)

Clearly, the transposition rule preserves the decisions and only alters their order. We, therefore, only need to show that the stochastic spine probabilities are preserved. For that, consider two reachable transposition-equivalent stochastic terms w_1 and w_2 . Referring to [SD.4], the probability of an arbitrary stochastic spine bag, in the right-hand side of the rule, going through branch i (of w_2), then through branch k (of w_1) is

$$r_i r_{ik} = \sum_{j=1}^{m_1} p_j p_{ji} \times \frac{p_k p_{ki}}{\sum_{j=1}^{m_1} p_j p_{ji}} = p_k p_{ki}$$

which is the probability of the same stochastic spine in the left-hand side of the rule.

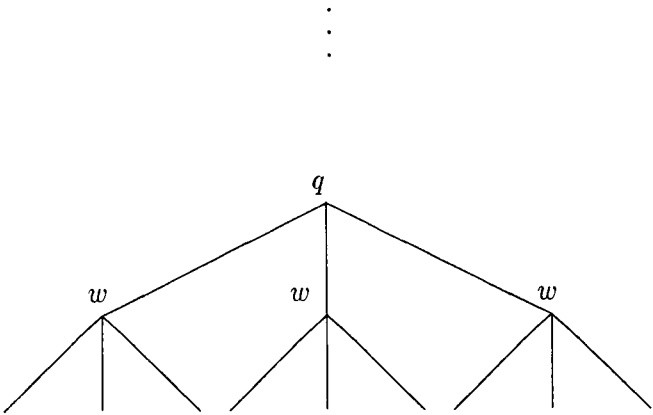
($\Rightarrow$)

Consider two stochastic spine set bag-equivalent stochastic terms. Consider the first decision in the first stochastic term. This decision occurs in every reachable stochastic spine of the first stochastic term. It therefore also occurs in every reachable stochastic spine of the second stochastic term. If this decision is not already at the root of the second stochastic term, it is a reachable semi-essential decision in this second term. In view of

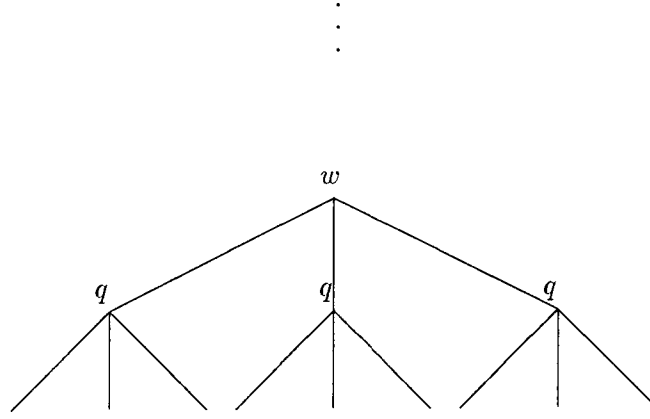
lemma 9, this decision can be moved to the root. As the terms are repeat consistent and repeat reduced the stochastic spine sets of the subterms are bag equivalent. Thus, repeat the procedure to the stochastic subterms of this common root until the second stochastic term has the same decision order as the first one. The transformed second stochastic term is completely identical to the first one. Thus the two stochastic terms are reachable transposition-equivalent.

B.4.3 Proof of Lemma 10 on Page 91

Suppose there is a series of manipulations, some of which are one-way, which leads to a stochastic term t' in which the query Q can be retrieved. We need to show that the one-way manipulations (idempotence) can actually be avoided in the process of obtaining the stochastic term t' and still obtain the same retrieval for the query. To this end, suppose that, at a given height in a two-way equivalent stochastic term of t , say t'' , there was a decision q which would have been eliminated by the use of the one-way manipulation of idempotence. Thus t'' must have the following form:



Regarding the subtree at the leaves of the decision q as a composite term, the transposition manipulation can be used to move q to the leaves of t'' as in the following diagram:



This has an effect which is equivalent to idempotence. As a result, the remaining manipulations can be carried out until the desired form in which the retrieval of the query $\mathcal{Q}$ is reached.

B.4.4 Proof of Lemma 11 on Page 92

($\Leftarrow$)

As $\mathcal{Q}'$ can be retrieved in t_{i_1} , there exists a stochastic term t'_{i_1} such that $t_{i_1} \Longleftrightarrow t'_{i_1}$ with t'_{i_1} having the required form for the query $\mathcal{Q}'$. We then have that

$$\begin{aligned} t \Longleftrightarrow t_0 &= q_1 : [a_{11}/p_{11} \gg t_1, \dots, a_{1i_1}/p_{1i_1} \gg t_{i_1}, \dots, a_{1n}/p_{1n} \gg t_n] \\ &\Longleftrightarrow t'_0 = q_1 : [a_{11}/p_{11} \gg t_1, \dots, a_{1i_1}/p_{1i_1} \gg t'_{i_1}, \dots, a_{1n}/p_{1n} \gg t_n]. \end{aligned}$$

Clearly, the stochastic term t'_0 has the required form for the query $\mathcal{Q}$. Thus, $\mathcal{Q}$ can be retrieved in t .

($\Rightarrow$)

As $\mathcal{Q}$ can be retrieved in t , there exists a stochastic term t' , which we may assume is repeat consistent and repeat reduced, such that $t \Longleftrightarrow t'$ with t' having the required form

for the query $\mathcal{Q}$. As

$$t \iff t_0 = q_1 : [a_{11}/p_{11} \gg t_1, \dots, a_{1i_1}/p_{1i_1} \gg t_{i_1}, \dots, a_{1n}/p_{1n} \gg t_n],$$

the stochastic decision $t' \iff t_0$. Clearly, the decision q_1 is semi-essential in t_0 , thus it must be reachable semi-essential in t' . As a result, there exists a stochastic term $t''_0 \iff t'$, rooted at q_1 and obtained by transposing the decision q_1 to the top of the tree. We then also have $t''_0 \iff t_0$ and both of these terms are rooted at the same decision q_1 . Clearly, t_{i_1} , repeat reduced if necessary, and the corresponding subtree in t''_0 are stochastic path set bag-equivalent, therefore, they are reachable transposition equivalent. As this subtree of t''_0 has the required form for the query $\mathcal{Q}'$, this query can be retrieved in the subtree t_{i_1} of t .

B.4.5 Proof of Theorem 11 on Page 92

($\implies$)

If $\mathcal{Q}$ can be retrieved in t_2 , there exists a stochastic term t'_2 such that $t_2 \iff t'_2$ with t'_2 having the required form for the query $\mathcal{Q}$. As $t_1 \implies t_2$, we have that $t_1 \implies t'_2$. This means that $\mathcal{Q}$ can be retrieved in t_1 .

($\impliedby$)

By induction on the height of t_2 . If the height of t_2 is 1, then

$$t_2 = q : [a_1/p_1 \gg \Omega, \dots, a_i/p_i \gg \Omega, \dots, a_m/p_m \gg \Omega]$$

Clearly, any query $\mathcal{Q}$ of the form $(q \langle ? \rangle a_i \mid \top)$ can be retrieved in t_2 . Therefore, by hypothesis, it can also be retrieved in t_1 . As a result, there exists a stochastic decision t'_1 such that $t_1 \iff t'_1$ with t'_1 rooted at q and with the same probabilities p_i 's on the branches of q . Clearly, $t'_1 \implies t_2$ by idempotence and thus $t_1 \implies t_2$.

Suppose that the result holds for stochastic terms of height less than or equal to $n - 1$.

Suppose that t_2 is of height n , then

$$t_2 = q : [a_1/p_1 \gg t_{21}, \dots, a_j/p_j \gg t_{2j}, \dots, a_m/p_m \gg t_{2m}]$$

with the t_{2i} 's having a height of at most $n - 1$. By hypothesis, every query Q which can be retrieved in t_2 can be retrieved in t_1 and $Ret_s(Q, t_1) = Ret_s(Q, t_2)$. In particular, consider the queries of the form $(q <?> a_j)$. Clearly, these can be retrieved in t_2 , thus they can also be retrieved in t_1 and $Ret_s(Q, t_1) = Ret_s(Q, t_2)$. As a result, there exists a stochastic term t'_1 , rooted at q and with the same probabilities p_i 's on the branches of q , such that $t_1 \iff t'_1$. Thus, t'_1 has the following form:

$$t'_1 = q : [a_1/p_1 \gg t'_{11}, \dots, a_j/p_j \gg t'_{1j}, \dots, a_m/p_m \gg t'_{1m}]$$

Consider now the arbitrary query whose conjunction contains the branch instruction $q <?> a_j$. If this query can be retrieved in t_2 , it can also be retrieved in t_1 and $Ret_s(Q, t_1) = Ret_s(Q, t_2)$. In view of Lemma 11, this query, diminished by the the branch instruction $q <?> a_j$, can be retrieved in t_{2j} and t'_{1j} . Since both of t_{2j} and t'_{1j} are of height $n - 1$, this implies that $t_{2j} \implies t'_{1j}$. As a result, we have that $t_1 \iff t'_1 \implies t_2$.

B.5 APPENDIX TO SECTION 4.6

B.5.1 Proof of Theorem 12 on Page 95

(i) The proof is by induction on the height of the stochastic term t :

If t has a zero height (*i.e.* a variable), the only retrievals which exist are of the form $Ret(\top, t) = 1$. Let M be a model satisfying t , it is clear that $Ret(\top, M) = 1$ and the result follows.

Suppose that the result holds for terms of height N . Consider the term

$$t = q : [a_1/p_1 \gg t_1, \dots, a_n/p_n \gg t_n,]$$

of height $N + 1$. If $C = \top$, then $Ret(C, t)$ exists, has value 1, and for every model satisfying t , $Ret(C, M) = 1$. If $C \neq \top$, we have either

1. $q<?>a_i \in C$

If $p_i = 0$, then $Ret(C, t)$ exists and has value 0. As

$$Ret(C, M) = \frac{m(M(C))}{m(\top_M)} \leq Ret(q<?>a_i, M) = \frac{m(M(q<?>a_i))}{m(\top_M)},$$

and $m(M(q<?>a_i)) = 0$ as M satisfies t , the result follows. If $p_i \neq 0$, then $Ret(C, t)$ exists and has value $p_i Ret(C', t_i)$, with $C' = C - \{q<?>a_i\}$, provided that $Ret(C', t_i)$ exists. Letting M be a model which satisfies t , and using the fact that $m(M(q<?>a_i)) = m(\top_{M_{q<\triangleright a_i}})$, we have that

$$\begin{aligned} Ret(C, M) &= \frac{m(M(C))}{m(\top_M)} \\ &= \frac{m(M(q<?>a_i) \cap M(C'))}{m(\top_M)} \\ &= \frac{m(M_{q<\triangleright a_i}(C'))}{m(\top_{M_{q<\triangleright a_i}})} \times \frac{m(M(q<?>a_i))}{m(\top_M)} \\ &= p_i Ret(C', t_i), \end{aligned}$$

as $m(M(\top_{M_{q<\triangleright a_i}})) \neq 0$, and by the induction hypothesis for the conjunction C' .

2. $q<?>a_i \notin C$

$$\begin{aligned}
 Ret(C, M) &= \frac{m(M_{q<\supset a_1}(C)) + \cdots + m(M_{q<\supset a_n}(C))}{m(\top_M)} \\
 &= p_1 \times \frac{m(M_{q<\supset a_1}(C))}{m(M(q<?>a_1))} + \cdots + p_n \times \frac{m(M_{q<\supset a_n}(C))}{m(M(q<?>a_n))} \\
 &= p_1 Ret(C, t_1) + \cdots + p_n Ret(C, t_n)
 \end{aligned}$$

The result then follows using the induction hypothesis for the conjunction C' .

- (ii) If $t_1 \implies t_2$, then the set of models satisfying t_1 is included in the set of models satisfying t_2 . As a result, any non-trivial model M which satisfies t_2 also satisfies t_1 . Since $Ret(C, t_2)$ exists, it is equal to $Ret(C, M)$. Thus, $Ret(C, t_1) = Ret(C, t_2) = Ret(C, M)$, provided that $Ret(C, t_1)$ exists.

Therefore, it remains only to show that $Ret(C, t_1)$ exists. To this end, we need to show that the existence of a retrieval is inherited through each application of [SD.1]–[SD.4] in the opposite direction to the inference. As t_1 is repeat-consistent, all its derivations are repeat-consistent. As a result, we only need to consider one single application between repeat-consistent terms. Furthermore, as the retrieval is recursively defined on subtrees, there will be no loss of generality if we assume that the rule is applied to the root of the term.

Among the four rules, only [SD.1] will be checked, as the result is quite obvious for the remaining rules. So, consider the rule

$$t_0 = q : [a_1/p_1 \gg t, \dots, a_n/p_n \gg t,] \implies t.$$

Suppose that $Ret(C, t)$ exists. If q does not occur in C , the result is clear. If q does occur in C , then in order for $q : [a_1/p_1 \gg t, \dots, a_n/p_n \gg t,]$ to be repeat-consistent, either q occurs in a reachable path of t and is certain along the same

branch at both occurrences, or q does not occur in any reachable path of t . In the former case, the inheritance of the retrieval is clearly insured. In the latter case, the only way in which $Ret(C, t)$ exists is the case where $Ret(C, t) = 0$, but then $Ret(C', t) = Ret(C - q \langle ? \rangle a_i, t) = 0$. Therefore, $Ret(C', t_0) = 0$, which implies that $Ret(C, t_0) = 0$ and thus exists.

- (iii) If $Ret(C, t)$ fails, there must be a non-empty set of leaves of the stochastic term t for which it is necessary to be able to perform a non-empty residual retrieval. The term t can be extended, at these leaves, in a repeat-consistent manner, by idempotently adding all the decisions from the condition of the residual retrieval with an arbitrary probability distribution. Each probability distribution defines a stochastic decision term and a different and disjoint set of models satisfying it. Consider the two particular stochastic decision terms, say t_1 and t_2 , defined by the probability distributions which assign, to a given element of the codomain of the added decisions, a probability of 1 and 0, respectively. Clearly, for any model M_1 satisfying t_1 and a model M_2 satisfying t_2 , The retrieval of C is, in general, different.

B.5.2 Proof of Lemma 12 on Page 96

($\Leftarrow$)

The structural retrieval provides a non-degenerate sequence of conjuncts on the path of the retrieval.

($\Rightarrow$)

By induction on the length of C . Suppose that it has length 0, that is $C = \top$, then for $Ret(q \langle ? \rangle a \mid \top, t)$ to exist the decision q must be reachable semi-essential in the stochastic term t and thus $Ret_s(q \langle ? \rangle a \mid \top, t)$ exists.

Suppose that the result holds for conjuncts of length N and C has length $N + 1$, then there exists a simple and non-degenerate conjunct C' , of length N , such that $C = C' \& q_0 \langle ? \rangle a_0$. By the induction hypothesis, $Ret_s(q_0 \langle ? \rangle a_0 \mid C', t)$ can be performed. This implies that there exists a term t' with a root path that corresponds to C such that $t' \iff t$. As $C \& q \langle ? \rangle a$ can be retrieved in t' , $q \langle ? \rangle a$ can be retrieved in the subtree below C . This means that the decision q is reachable semi-essential in this subtree implying that the structural retrieval $Ret_s(q \langle ? \rangle a \mid C, t)$ can be performed.

B.5.3 Proof of Theorem 13 on Page 97

($\implies$)

By soundness.

($\impliedby$)

As $M(t_1) \subseteq M(t_2)$, we have that $\mathcal{C}_{M(t_2)} \subseteq \mathcal{C}_{M(t_1)}$. Every non-degenerate simple conjunct in $\mathcal{C}_{M(t_2)}$ transmits by containment to one in $\mathcal{C}_{M(t_1)}$. This means that every structural retrieval in t_2 can be performed in t_1 and thus $t_1 \implies t_2$.

APPENDIX C

APPENDIX TO CHAPTER 5

C.1 APPENDIX TO SECTION 5.1

C.1.1 Proof of Lemma 13 on Page 99

If $t = x$, a variable, then t is certain and $H(t) = 0$. If t is not a variable, then it is a stochastic term of the form $t = q : [a_1/p_1 \gg t_1, \dots, a_n/p_n \gg t_n]$, in which the subterm t_i has n_i paths each of which, say the j^{th} path, has a probability of occurrence of r_{ij} . As a result, the entropy $H(t_i)$ of the stochastic subterm t_i is by definition $H(t_i) = -\sum_{j=1}^{n_i} r_{ij} \log r_{ij}$. The entropy $H(t)$ of the stochastic term t can then be computed as follows:

$$\begin{aligned} H(t) &= -\sum_{i=1}^n \sum_{j=1}^{n_i} p_i r_{ij} \log(p_i r_{ij}) = -\sum_{i=1}^n p_i \log p_i \sum_{j=1}^{n_i} r_{ij} - \sum_{i=1}^n p_i \sum_{j=1}^{n_i} r_{ij} \log r_{ij} \\ &= -\sum_{i=1}^n p_i \log p_i - \sum_{i=1}^n p_i H(t_i). \end{aligned}$$

C.1.2 Proof of Lemma 14 on Page 100

As the inference rule [SD.5] results from a combination of [SD.1] and [SD.2], we only need to check the lemma for the inference rules [SD.1]–[SD.4].

(1) [SD.1]: If $t_1 \implies t_2$ by idempotence, then there exists a decision q such that

$$t_1 = q : [a_1/p_1 \gg t_2, \dots, a_m/p_m \gg t_2].$$

In view of Lemma 13,

$$H(t_1) = -\sum_{i=1}^m p_i \log p_i + \sum_{i=1}^m p_i H(t_2) = -\sum_{i=1}^m p_i \log p_i + H(t_2).$$

Obviously, $H(t_1) \geq H(t_2)$.

(2) [SD.2]: Obvious as, in view of Lemma 13, terms occurring on a zero-probability path do not contribute entropy.

(3) [SD.3]: If $t_1 \Longleftrightarrow t_2$ by repetition, then there exists a decision q such that

$$t_1 = q : [a_1/p_1 \gg x_1, \dots, a_i/p_i \gg x_i, \dots, a_m/p_m \gg x_m]$$

and

$$t_2 = q : [a_1/p_1 \gg x_1, \dots, a_i/p_i \gg t_0, \dots, a_m/p_m \gg x_m],$$

where

$$t_0 = q : [a_1/0 \gg x_1, \dots, a_i/1 \gg x_i, \dots, a_m/0 \gg x_m].$$

Using Lemma 13,

$$H(t_2) = -\sum_{i=1}^m (p_i \log p_i) + p_i H(t_0) = H(t_1),$$

as $H(t_0) = 0$.

(4) [SD.4]: The transposition manipulation conserves the probability of all the paths and simply alters their order. As a result, the entropy remains unchanged after a transposition manipulation.

C.1.3 Proof of Corollary 2 on Page 100

As $t_1 \Longleftrightarrow t_2$, $t_1 \Rightarrow t_2$ and $t_2 \Rightarrow t_1$. This implies that $H(t_1) \geq H(t_2)$ and $H(t_2) \geq H(t_1)$, respectively. Thus, $H(t_1) = H(t_2)$.

C.2 APPENDIX TO SECTION 5.2

C.2.1 Proof of Theorem 14 on Page 102

Consider a stochastic term t in which a number of stochastic terms, $t_{l_1}, t_{l_2}, \dots, t_{l_M}$, not necessarily distinct, are missing on M given paths $P_{l_1}, P_{l_2}, \dots, P_{l_M}$ among the N paths $P_1, P_2, \dots, P_N$ of the stochastic term t . Let $p_1, p_2, \dots, p_N$ be the probability of occurrence of the paths of the stochastic term t . The entropy $H(t)$ of the stochastic term t is $H(t) = -\sum_{i=1}^N p_i \log p_i$. In order to complete the stochastic term t with the missing information (the stochastic terms t_{l_j} 's), append each t_{l_j} to the leaf of the corresponding path P_{l_j} where it is missing. Call the resulting stochastic term t' . Clearly, according to Lemma 13, the entropy $H(t')$ of the stochastic term t' is

$$H(t') = H(t) + \sum_{j=1}^M p_{l_j} H(t_{l_j}).$$

Let $r_{j1}, r_{j2}, \dots, r_{jn_j}$ be the unknown probability of occurrence of the paths of t_{l_j} . The entropy of each stochastic term t_{l_j} is by definition $H(t_{l_j}) = -\sum_{k=1}^{n_j} r_{jk} \log r_{jk}$. As a result, the entropy $H(t')$ of the stochastic term t' becomes

$$H(t') = H(t) - \sum_{j=1}^M p_{l_j} \sum_{k=1}^{n_j} r_{jk} \log r_{jk}.$$

We now seek a probability distribution of the r_{jk} 's which maximizes the entropy $H(t')$ of the stochastic term t' . For that, we only need to maximize $\sum_{j=1}^M p_{l_j} H(t_{l_j})$ with respect to the r_{jk} 's, which are constrained by $\sum_k^{n_j} r_{jk} = 1$.

Let $n\lambda = \begin{bmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_M \end{bmatrix}^T$ be a vector of Lagrange multipliers which solves

$$\begin{cases} \frac{\partial \sum_{j=1}^M p_{l_j} H(t_{l_j})}{\partial r_{jk}} + \lambda_j \times \frac{\partial (1 - \sum_{k=1}^{n_j} r_{jk})}{\partial r_{jk}} = 0 \\ 1 - \sum_{k=1}^{n_j} r_{jk} = 0, \end{cases}$$

for $j = 1, \dots, M$. The first equation leads to $r_{jk} = \exp(1 - \frac{\lambda_j}{p_{l_j}})$. The second equation yields that $\lambda_j = p_{l_j}(1 + \log n_j)$, which implies that $r_{jk} = \frac{1}{n_j}$.

C.2.2 Proof of Theorem 15 on Page 103

Consider a stochastic term t in which a number of stochastic terms, $t_{l_1}, t_{l_2}, \dots, t_{l_M}$, not necessarily distinct, are missing on M given paths $P_{l_1}, P_{l_2}, \dots, P_{l_M}$ among the N paths $P_1, P_2, \dots, P_N$ of the stochastic term t . Let $p_1, p_2, \dots, p_N$ be the probability of occurrence of the paths of the stochastic term t . The entropy $H(t)$ of the stochastic term t is $H(t) = -\sum_{i=1}^N p_i \log p_i$. In order to complete the stochastic term t with the missing information (the stochastic terms t_{l_j} 's), append each t_{l_j} to the leaf of the corresponding path P_{l_j} where it is missing. Call the resulting stochastic term t' . Clearly, according to Lemma 13, the entropy $H(t')$ of the stochastic term t' is

$$H(t') = H(t) + \sum_{j=1}^M p_{l_j} H(t_{l_j}).$$

Let $r_{j1}, r_{j2}, \dots, r_{jn_j}$ be the unknown probability of occurrence of the paths of t_{l_j} . The entropy of each stochastic term t_{l_j} is by definition $H(t_{l_j}) = -\sum_{k=1}^{n_j} r_{jk} \log r_{jk}$. As a result, the entropy $H(t')$ of the stochastic term t' becomes

$$H(t') = H(t) - \sum_{j=1}^M p_{l_j} \sum_{k=1}^{n_j} r_{jk} \log r_{jk}.$$

We now seek a probability distribution of the r_{jk} 's which maximizes the entropy $H(t')$ of the stochastic term t' . For that, we only need to maximize $\sum_{j=1}^M p_{l_j} H(t_{l_j})$ with respect to the r_{jk} 's, which are constrained by $\sum_{k=1}^{n_j} r_{jk} = 1$. Suppose that additional m linear inequalities also constrain the r_{jk} 's. These constraints may take the form of

$$\begin{cases} a_{11}r_{j1} + \cdots + a_{1n_j}r_{jn_j} & \leq 0, \\ & \vdots \\ a_{m1}r_{j1} + \cdots + a_{mn_j}r_{jn_j} & \leq 0, \end{cases}$$

To recapitulate, the problem of probabilistic entailment in the presence of linear inequalities is equivalent to solving the following set of nonlinear programs for $j = 1, \dots, M$.

$$\begin{array}{ll} \text{minimize} & r_{j1} \log r_{j1} + \cdots + r_{jn_j} \log r_{jn_j} \\ \text{subject to} & \begin{cases} a_{11}r_{j1} + \cdots + a_{1n_j}r_{jn_j} \leq 0, \\ \vdots \\ a_{m1}r_{j1} + \cdots + a_{mn_j}r_{jn_j} \leq 0, \\ \sum_{k=1}^{n_j} r_{jk} = 1, \\ r_{j1} \geq 0, \dots, r_{jn_j} \geq 0. \end{cases} \end{array}$$

APPENDIX D

APPENDIX TO CHAPTER 6

D.1 APPENDIX TO SECTION 6.1

D.1.1 Proof of Theorem 16 on Page 109

The proof consists of the following cyclic implications $(i) \implies (ii) \implies (iii) \implies (i)$.

$((i) \implies (ii))$

Given that the l decisions are mutually statistically independent, any ordering of these decisions satisfies the above $l - 1$ conditional probability equations. The following development shows that any one of these equations (say the $(k - 1)^{th}$) indeed holds.

$$\begin{aligned}
 & p(q_{l_k} \langle ? \rangle a_{l_k i_k} \mid q_{l_{k-1}} \langle ? \rangle a_{l_{k-1} i_{k-1}}, \dots, q_{l_2} \langle ? \rangle a_{l_2 i_2}, q_{l_1} \langle ? \rangle a_{l_1 i_1}) \\
 &= \frac{p(q_{l_k} \langle ? \rangle a_{l_k i_k}, q_{l_{k-1}} \langle ? \rangle a_{l_{k-1} i_{k-1}}, \dots, q_{l_2} \langle ? \rangle a_{l_2 i_2}, q_{l_1} \langle ? \rangle a_{l_1 i_1})}{p(q_{l_{k-1}} \langle ? \rangle a_{l_{k-1} i_{k-1}}, \dots, q_{l_2} \langle ? \rangle a_{l_2 i_2}, q_{l_1} \langle ? \rangle a_{l_1 i_1})} \\
 &= \frac{p(q_{l_k} \langle ? \rangle a_{l_k i_k}) p(q_{l_{k-1}} \langle ? \rangle a_{l_{k-1} i_{k-1}}) \cdots p(q_{l_2} \langle ? \rangle a_{l_2 i_2}) p(q_{l_1} \langle ? \rangle a_{l_1 i_1})}{p(q_{l_{k-1}} \langle ? \rangle a_{l_{k-1} i_{k-1}}) \cdots p(q_{l_2} \langle ? \rangle a_{l_2 i_2}) p(q_{l_1} \langle ? \rangle a_{l_1 i_1})} \\
 &= p(q_{l_k} \langle ? \rangle a_{l_k i_k}).
 \end{aligned}$$

Note that if one or more of the pre-conditions in this conditional probability has a zero probability, thus making the conditional probability undefined, this conditional probability can be arbitrarily set, by the use of the indifference rule [SD.3] to satisfy the desired condition. This is equivalent to say that events occurring on a zero-probability path can be simply simply ignored, which makes sense.

((ii) $\implies$ (iii))

Obvious.

((iii) $\implies$ (i))

Given that we have an ordering $l_1, l_2, \dots, l_l$ of the l decisions $q_1, q_2, \dots, q_l$, for which we have

$$\begin{aligned} p(q_{l_2} \langle ? \rangle a_{l_2 i_2} \mid q_{l_1} \langle ? \rangle a_{l_1 i_1}) &= p(q_{l_2} \langle ? \rangle a_{l_2 i_2}), \\ p(q_{l_3} \langle ? \rangle a_{l_3 i_3} \mid q_{l_2} \langle ? \rangle a_{l_2 i_2}, q_{l_1} \langle ? \rangle a_{l_1 i_1}) &= p(q_{l_3} \langle ? \rangle a_{l_3 i_3}), \\ &\dots \\ p(q_{l_l} \langle ? \rangle a_{l_l i_l} \mid q_{l_{l-1}} \langle ? \rangle a_{l_{l-1} i_{l-1}}, \dots, q_{l_2} \langle ? \rangle a_{l_2 i_2}, q_{l_1} \langle ? \rangle a_{l_1 i_1}) &= p(q_{l_l} \langle ? \rangle a_{l_l i_l}), \\ \forall i_j \in \{1, \dots, m_{l_j}\} \text{ and } \forall j \in \{1, \dots, l\}, \end{aligned}$$

we need to show that $\forall L_1, L_2, \dots, L_{L_0} \leq l$, we have that

$$\begin{aligned} &p(q_{L_1} \langle ? \rangle a_{L_1 i_1}, q_{L_2} \langle ? \rangle a_{L_2 i_2}, \dots, q_{L_{L_0}} \langle ? \rangle a_{L_{L_0} i_{L_0}}) \\ &= \\ &p(q_{L_1} \langle ? \rangle a_{L_1 i_1}) p(q_{L_2} \langle ? \rangle a_{L_2 i_2}) \cdots p(q_{L_{L_0}} \langle ? \rangle a_{L_{L_0} i_{L_0}}), \\ &\forall i_j \in \{1, \dots, m_{L_j}\} \text{ and } \forall j \in \{1, \dots, L_0\}. \end{aligned}$$

We now take the joint probability distribution

$$p(q_{L_1} \langle ? \rangle a_{L_1 i_1}, q_{L_2} \langle ? \rangle a_{L_2 i_2}, \dots, q_{L_{L_0}} \langle ? \rangle a_{L_{L_0} i_{L_0}}),$$

in which the decision indices L_k 's match only $L_0 \leq l$ indices of the ordering $l_1, l_2, \dots, l_l$.

Rewrite the joint probability distribution to include the decisions of the un-matched indices by summing over the codomains of the un-matched decisions as follows:

$$\begin{aligned} &p(q_{L_1} \langle ? \rangle a_{L_1 i_1}, q_{L_2} \langle ? \rangle a_{L_2 i_2}, \dots, q_{L_{L_0}} \langle ? \rangle a_{L_{L_0} i_{L_0}}) \\ &= \\ &\underbrace{\sum \cdots \sum}_{l-L_0 \text{ times}} p(q_{l_1} \langle ? \rangle a_{l_1 i_1}, q_{l_2} \langle ? \rangle a_{l_2 i_2}, \dots, q_{l_l} \langle ? \rangle a_{l_l i_l}). \end{aligned}$$

In view of this result, using the chain rule formula [56], we obtain that

$$\begin{aligned}
& \underbrace{\sum \cdots \sum}_{l-L_0 \text{ times}} p(q_{l_1} \langle ? \rangle a_{l_1 i_1}, q_{l_2} \langle ? \rangle a_{l_2 i_2}, \dots, q_{l_l} \langle ? \rangle a_{l_l i_l}) \\
& = \\
& \underbrace{\sum \cdots \sum}_{l-L_0 \text{ times}} p(q_{l_l} \langle ? \rangle a_{l_l i_l} \mid q_{l_{l-1}} \langle ? \rangle a_{l_{l-1} i_{l-1}}, \dots, q_{l_2} \langle ? \rangle a_{l_2 i_2}, q_{l_1} \langle ? \rangle a_{l_1 i_1}) \cdots \\
& \cdots p(q_{l_3} \langle ? \rangle a_{l_3 i_3} \mid q_{l_2} \langle ? \rangle a_{l_2 i_2}, q_{l_1} \langle ? \rangle a_{l_1 i_1}) p(q_{l_2} \langle ? \rangle a_{l_2 i_2} \mid q_{l_1} \langle ? \rangle a_{l_1 i_1}) p(q_{l_1} \langle ? \rangle a_{l_1 i_1}) \\
& = \\
& \underbrace{\sum \cdots \sum}_{l-L_0 \text{ times}} p(q_{l_l} \langle ? \rangle a_{l_l i_l}) \cdots p(q_{l_3} \langle ? \rangle a_{l_3 i_3}) p(q_{l_2} \langle ? \rangle a_{l_2 i_2}) p(q_{l_1} \langle ? \rangle a_{l_1 i_1}) \\
& = \\
& p(q_{L_1} \langle ? \rangle a_{L_1 i_1}) p(q_{L_2} \langle ? \rangle a_{L_2 i_2}) \cdots p(q_{L_{L_0}} \langle ? \rangle a_{L_{L_0} i_{L_0}}).
\end{aligned}$$

D.1.2 Proof of Corollary 3 on Page 110

The necessary and sufficient conditions for mutual statistical independence require that the conditional probabilities of the decisions (depicted in some order) be equal, in all the paths of the stochastic term, to the *a priori* probabilities of these decisions. Clearly, all the decisions must be reachable semi-essential for these requirements to hold.

D.2 APPENDIX TO SECTION 6.2

D.2.1 Proof of Lemma 15 on Page 111

- (i) We first need to prove that $\mathcal{P}(S) \subset \mathcal{P}_{a_1}(S) \cup \mathcal{P}(S_1)$. For that, let $X \in \mathcal{P}(S)$; then we either have $a_1 \in X$ or $a_1 \notin X$. If $a_1 \in X$, then by definition $X \in \mathcal{P}_{a_1}(S)$. If $a_1 \notin X$, then $X \in \mathcal{P}(S_1)$.

(ii) We then need to prove that $\mathcal{P}_{a_1}(S) \cup \mathcal{P}(S_1) \subset \mathcal{P}(S)$. Consider $X \in \mathcal{P}_{a_1}(S) \cup \mathcal{P}(S_1)$, then we either have $X \in \mathcal{P}_{a_1}(S)$ or $X \in \mathcal{P}(S_1)$. If $X \in \mathcal{P}_{a_1}(S)$, then by definition $X \in \mathcal{P}(S)$. If $X \in \mathcal{P}(S_1)$, then $X \in \mathcal{P}(S)$ as $S_1 \subset S$.

APPENDIX E

APPENDIX TO CHAPTER 7

E.1 APPENDIX TO SECTION 7.3

E.1.1 Proof of Theorem 17 on Page 125

The proof is by induction on the number of observations in the sequence. Let t_0 be the stochastic decision tree describing the system before any observation is received. This stochastic decision tree will feature the decisions partitioning the state space of the system and give the *a priori* probability of the states, namely $p(x_i)$, $\forall i = 1, \dots, N$. Suppose that k observations have been processed and that the system given these observations is described by the stochastic decision tree t_k . The decision tree t_k can be encapsulated as a composite term (see Chapter 4 or [15]), and thus can be written as

$$t_k = w : [x_1/p(x_1 | Y_k), \dots, x_N/p(x_N | Y_k)].$$

Suppose now that the observation y_{k+1} is received and the conditional probabilities $p(y_{k+1} | x_i)$, $\forall i = 1, \dots, N$ are available. This observation can be viewed as an additional decision $\mathcal{Y}_{k+1}$, the observation decision, with only one element of the codomain (y_{k+1}), which is appended at the leaves of t_k (the system states) with the corresponding conditional probabilities. A stochastic decision t'_k is obtained and is

$$t'_k = w : [x_1/p(x_1 | Y_k) \gg \mathcal{Y}_{k+1} : [y_{k+1}/p(y_{k+1} | x_1)], \dots, \\ \dots, x_N/p(x_N | Y_k) \gg \mathcal{Y}_{k+1} : [y_{k+1}/p(y_{k+1} | x_N)]] .$$

Moving $\mathcal{Y}_{k+1}$ to the top of the tree t'_k by the use of the transposition manipulation ([SD.4]) for composite terms, an equivalent (to t'_k) stochastic decision t'_{k+1} is obtained, with

$$t'_{k+1} = \mathcal{Y} : [y_{k+1}/p \gg w : [x_1/p_1, \dots, x_N/p_N]],$$

with

$$\begin{cases} p &= \sum_{i=1}^n p(y_{k+1} | x_i) \times p(x_i | Y_k) &= p(y_{k+1} | Y_k) \\ p_i &= \frac{p(y_{k+1} | x_i) \times p(x_i | Y_k)}{\sum_{i=1}^n p(y_{k+1} | x_i) \times p(x_i | Y_k)}. \end{cases}$$

Clearly, the conditional probabilities p_i , $\forall i = 1, \dots, N$, labeling the branches of the subtree t_{k+1} of t'_{k+1} , rooted at w , coincide with the estimator updated conditional probability after the observation y_{k+1} is processed. The induction is completed by noting that t_0 is the decision tree defined by the initial probability distribution $p(x_i)$, $i = 1, \dots, N$.

E.2 APPENDIX TO SECTION 7.4

E.2.1 Proof of Lemma 18 on Page 132

We need to show that the relation $\mathcal{R}_0$ is reflexive, symmetric, and transitive.

- (i) Reflexivity: follows from the reflexivity of the binary relation $\mathcal{R}$.
- (ii) Symmetry: follows from the symmetry of the binary relation $\mathcal{R}$ and from the symmetry of the expression in the definition of $\mathcal{R}_0$ involving the sequence of decisions.
- (iii) Transitivity: Suppose that for $q_1, q_2, q_3 \in S$, we have that $q_1 \mathcal{R}_0 q_2$ and $q_2 \mathcal{R}_0 q_3$. We need to show that $q_1 \mathcal{R}_0 q_3$ as well.

There are four possibilities in the hypothesis; they are:

1. $q_1 \mathcal{R} q_2$ and $q_2 \mathcal{R} q_3$.
2. $q_1 \mathcal{R} q_2$ and there exists a sequence of decisions in S , $\{q_a, q_b, \dots, q_n\}$, such that $q_2 \mathcal{R} q_a, q_a \mathcal{R} q_b, \dots$, and $q_n \mathcal{R} q_3$.
3. There exists a sequence of decisions in S , $\{q_a, q_b, \dots, q_n\}$, such that $q_1 \mathcal{R} q_a, q_a \mathcal{R} q_b, \dots$, and $q_n \mathcal{R} q_2$ and $q_2 \mathcal{R} q_3$.
4. There exists a sequence of decisions in S , $\{q_{a1}, q_{b1}, \dots, q_{n1}\}$, such that $q_1 \mathcal{R} q_{a1}, q_{a1} \mathcal{R} q_{b1}, \dots$, and $q_{n1} \mathcal{R} q_2$ and there exists a sequence of decisions in S , $\{q_{a2}, q_{b2}, \dots, q_{n2}\}$, such that $q_2 \mathcal{R} q_{a2}, q_{a2} \mathcal{R} q_{b2}, \dots$, and $q_{n2} \mathcal{R} q_3$.

We have to prove transitivity ($q_1 \mathcal{R}_0 q_3$) in each one of these four cases.

Case 1: $\{q_2\}$ is the sequence connecting q_1 to q_3 . Thus, by definition, $q_1 \mathcal{R}_0 q_3$.

Case 2: $\{q_2, q_a, q_b, \dots, q_n\}$ is the sequence connecting q_1 to q_3 . Thus, by definition, $q_1 \mathcal{R}_0 q_3$.

Case 3: $\{q_a, q_b, \dots, q_n, q_2\}$ is the sequence connecting q_1 to q_3 . Thus, by definition, $q_1 \mathcal{R}_0 q_3$.

Case 4: $\{q_{a1}, q_{b1}, \dots, q_{n1}, q_2, q_{a2}, q_{b2}, \dots, q_{n2}\}$ is the sequence connecting q_1 to q_3 . Thus, by definition, $q_1 \mathcal{R}_0 q_3$.

E.2.2 Proof of Lemma 19 on Page 133

The proof is established by the two-way inclusion:

1. $\mathcal{P} \subset D_f$: If $X \in \mathcal{P}$, then the decisions in X are either dependent on each other or connected to each other by a sequence of dependent decisions. This means that

$X \in D$ or $X = \cup_i X_i$, with the X_i 's being at least pairwise non-disjoint. This implies that $X \in D_f$.

2. $\mathcal{P} \supset D_f$: If $X \in D_f$, then either $X \in D$, in which case X must belong to $\mathcal{P}$, or $X = \cup_i X_i$, with the X_i 's being at least pairwise non-disjoint. This also means that $X \in \mathcal{P}$.

E.2.3 Proof of Theorem 18 on Page 138

Given a set of lower order probability distributions $P_1, P_2, \dots, P_m$, such that their product $P' = P_1 P_2 \cdots P_m$ is a product approximation to their higher probability distribution, we need to show that of the entire class of common extensions of these lower order components, P' has maximum entropy.

To this end, consider an arbitrary common extension of $P_1, P_2, \dots$, and P_m ; call it P'' . According to a well-known result in information theory (see for example [25]),

$$\sum_j P''_j \log P''_j \geq \sum_j P'_j \log P'_j.$$

The right-hand side of this inequalities can be developed to

$$\begin{aligned} \sum_j P''_j \log P''_j &= \sum_j P''_j \log(P_1 P_2 \cdots P_m) \\ &= \sum_j P''_j \log P_1 + \sum_j P''_j \log P_2 + \cdots + \sum_j P''_j \log P_m. \end{aligned}$$

Since P'' is also an extension of $P_1, P_2, \dots$, and P_m , the right-hand side of the above equalities can be properly summed to obtain that

$$\sum_j P''_j \log P''_j = \sum P_1 \log P_1 + \sum P_2 \log P_2 + \cdots + \sum P_m \log P_m = \sum_j P'_j \log P'_j.$$

Thus, the original expression becomes $\sum_j P''_j \log P''_j \geq \sum_j P'_j \log P'_j$, which is the desired result.

VITA

Jalel Zrida was born in Sfax, Tunisia, on March 22, 1956. In June 1975, he received the *Baccalaureat* in Mathematics from the *Lycée de Garçons de Sfax, Tunisie*. He then entered the *Faculté de Sciences et Techniques de Sfax, Tunisie*, where he received, in June 1981, the *Diplôme d'Ingénieur Principal* in Electrical Engineering and Instrumentation. In Fall 1981, he entered the Department of Electrical & Computer Engineering at the University of Tennessee, Knoxville, to pursue a Ph.D degree in Electrical Engineering in the Control Systems area. During his studies, Mr. Zrida was employed, at the Electrical & Computer Engineering Department, as a teaching assistant, a research assistant, and an instructor. He also worked at the Oak Ridge National Laboratory at Oak Ridge, Tennessee, as a research assistant in the Energy Division. His research includes development of residential load models for direct load control, expert systems applications, and stochastic decision theory.

Mr. Zrida is a Sigma Xi full member, a SPIE member, and an IEEE member.