Algebraic Uncertainty Theory

A Unifying Perspective on Reasoning under Uncertainty

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Jörg Zimmermann

aus

Frankfurt am Main

Bonn, 2012

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn.

1. Gutachter: Univ.-Prof. Dr. Armin B. Cremers

2. Gutachter: Univ.-Prof. Dr. Stefan Wrobel

Tag der Promotion: 30. August 2012

Erscheinungsjahr: 2013

Abstract

The question of how to represent and process uncertainty is of fundamental importance to the scientific process, but also in everyday life. Currently there exist a lot of different calculi for managing uncertainty, each having its own advantages and disadvantages. Especially, almost all are defining the domain and structure of uncertainty values a priori, e.g., one real number, two real numbers, a finite domain, and so on, but maybe uncertainty is best measured by complex numbers, matrices or still another mathematical structure. This thesis investigates the notion of uncertainty from a foundational point of view, provides an ontology and axiomatic core system for uncertainty and derives and not defines the structure of uncertainty. The main result, the ring theorem, stating that uncertainty values are elements of the [0,1]-interval of a partially ordered ring, is used to derive a general decomposition theorem for uncertainty values, splitting them into a numerical interval and an "interaction term". In order to illustrate the unifying power of these results, the relationship to Dempster-Shafer theory is discussed and it is shown that all Dempster-Shafer measures over finite domains can be represented by ring-valued uncertainty measures. Finally, the historical development of approaches to modeling uncertainty which have led to the results of this thesis are reviewed.

Acknowledgements

An acknowledgement is a reflection on an important precondition of successful projects: a social network of people who directly or indirectly contribute to these projects in various ways. This thesis is also embedded in such a "web of life", connecting the efforts of many individuals and thereby transcending the limits of time and space. Here I can only express my deeply felt gratitude to the most directly involved persons, and can only offer my apologies to those who I have forgotten to mention.

Prof. Dr. Armin B. Cremers has given me invaluable advice, support and feedback over many years. He helped me to keep focus and reorder my priorities when the number of possibilities to explore became overwhelming. He also managed to create an atmosphere where free exploration of ideas and constructive criticism could coexist, which helped to establish a productive working environment, an environment which I learned to appreciate more and more during our collaboration and which was essential for the completion of this thesis.

I want to thank Prof. Dr. Stefan Wrobel, Prof. Dr. Andreas Weber, and Prof. Dr. Steve Horvath for agreeing to be referees of this thesis and investing their time and considerations. Their encouragement and enthusiasm have greatly contributed to keeping up my motivation and staying on track.

My close friends and colleagues Henning Henze, Torsten Nahm, Uwe Radetzki and Martin Kutz are part of this thesis in many ways, direct and indirect, conscious and unconscious, helping me with advice, feedback, and ideas or by just being there when I needed them. Tragically, Martin Kutz has died during the work on this thesis. I will miss the extremely lucid and fruitful discussions I had with him, this collaborative development of half-baked ideas which is so invaluable in clearing one's mind. Martin, I will always remember you.

Finally, my parents and family have filled my cup of love in abundance, which has freed my mind to pursue the higher levels of the Maslow pyramid. Without their endless encouragement, patience, and support this thesis would not exist.

Contents

1 Introduction **6**

2 Analysis of Uncertainty **8**

2.1 An Ontology of Uncertainty 9

2.2 Indefiniteness 9

2.3 Types of Uncertainty 10

2.3.1 Event Uncertainty: 10

2.3.2 Model Uncertainty: 11

2.3.3 Severe Uncertainty: 11

2.4 Existing Approaches to Uncertainty 11

3 Formalizing Uncertainty **12**

3.1 The Algebra of Truth Bearers 13

3.2 Uncertainty: the Boolean Case 13

3.3 Axioms for Uncertainty Measures 14

4 Structure Theorems for Uncertainty **16**

4.1 Basic Properties of Confidence Measures 17

4.2 Extensions of Partially Ordered Structures 18

4.3 Proper Confidence Structures 24

4.4 Ring Theorem 29

4.4.1 Extension of G into a total function 29

4.4.2 Cancellation property of G 31

4.4.3 Semiring property of G and F 32

4.4.4 Extension of G and F into a ring 33

4.5 Probability theory as a model of NC_{12} 33

4.6 Total Order Theorem 35

5 Uncertainty Decomposition **37**

6 The Lineage of NC_{12} **41**

6.1 The Axiom System of Cox 43

6.1.1 The counterexample of Halpern 43

6.2 The Axiom System of Paris 45

6.3 The Axiom System of Arnborg and Sjödin 46

7 Relations to existing Uncertainty Calculi **47**

7.1 Lower Probabilities 47

7.2 Dempster-Shafer Theory 49

7.2.1 Interference Families 51

7.3 Non-monotonic Logic 52

8 Conclusions **53**

9 Effective Learning Systems **54**

9.1 Algorithmic Ontology: Programs as Generators 54

9.2 Learning Systems 55

9.3 Synchronous Learning Framework 57

9.4 Conclusions 58

References **60**

1 Introduction

Nothing is more important than to see the sources of invention, which are, in my opinion, more interesting than the inventions themselves.

Gottfried Wilhelm Leibniz

The starting point of this thesis was the following question: "What are the fundamental possibilities and limitations of learning by an *effective* system?"

An effective learning system is a system which can be fully specified by a program on a universal Turing machine. In its most general form, this program transforms a stream of percepts generated by an environment into a stream of actions possibly changing this environment. Via this senso-motoric loop the system is embedded into its environment, about which a priori nothing is known. A more specific notion of learning is defined as the process which translates the stream of percepts into predictions for future percepts. These predictions can then be used for choosing actions. The analysis of the "design space" for effective learning systems leads to three major questions:

1. How should a learning system represent and process uncertainty, or, what is the proper inductive logic?
2. What set of possible models of the environment should the system consider?
3. How to relate the explanatory power of a model to its complexity?

In the long run, the learning system should be able to detect as many regularities in its percept stream as possible, while dealing sensibly with the inherent uncertainty of predictions based on a finite amount of data.

It has turned out that already the first question was enough of a quagmire to absorb much of the work of several years, so this thesis will focus on this question. However, there will be a summary of ideas and preliminary results with regard to the last two questions.

The main contribution of this thesis can be summarized as follows:

A *minimalistic axiom system* for uncertainty measures (also referred to as confidence theory) is introduced which entails a ring structure for uncertainty values. Classical probability theory is a special case and other uncertainty calculi,

like Dempster-Shafer theory, can be related to and interpreted within this axiom system, thereby providing a unifying perspective on the multitude of approaches developed for modeling uncertainty.

2 Analysis of Uncertainty

The quest for a theory of inductive logic, i.e., a logic defining the relationship between observations and hypotheses, lies at the heart of the scientific process. Accordingly, there is a plethora of research aiming at the clarification of this relationship, which has led to a lot of different calculi for managing uncertainty, each having its own advantages and disadvantages. Especially, almost all are defininining the domain and structure of uncertainty values a priori, e.g., one real number, two real numbers, a finite domain, and so on, but maybe uncertainty is best measured by complex numbers, matrices or still another mathematical structure. This thesis introduces an approach which leaves the domain of uncertainty values a priori undefined (an unstructured set) and derives its algebraic structure from axioms concerning only general properties of uncertainty measures. This approach to uncertainty calculi can be denoted as *algebraic uncertainty theory* and we think that such a framework is well-suited to investigate the commonalities and differences of existing uncertainty calculi and to provide a reference system for general results in this area of research. First results in this direction are published in [ZC11].

The introduced axioms are based on a recent axiomatization for uncertainty measures given by Arnborg and Sjödin [AS01]. Their axiom system is in the line of thinking started by R. T. Cox in 1946, and removes one important obstacle to a widespread acceptance of Cox's system: the assumption that uncertainty values should be measured by real numbers, excluding approaches like Dempster-Shafer theory or other, more complex structured domains of uncertainty values. Their main result is that, given their axiom system, the domain of uncertainty values has field structure and the classical axioms of probability theory hold with regard to the operations of this field. However, in a later part of their analysis, Arnborg and Sjödin introduce a total order assumption for the domain of uncertainty values, thus restricting the scope of their result. Here we give a reworked and streamlined version of their axiom system, most importantly dropping the total order assumption and an axiom concerning disjunction of propositions. Without total order assumption, in general a domain of uncertainty values exhibits only ring structure instead of field structure, as there are now domains containing zero divisors. Furthermore, a weak additional assumption on the order structure of uncertainty values, i.e. that they are lattice-ordered, implies that the domain of uncertainty values is totally ordered, giving rise to the following trilemma: the three properties of partial order, lattice-order, and field structure of a domain of uncertainty values cannot be satisfied all at the same time.

2.1 An Ontology of Uncertainty

In the realm of empirical knowledge, uncertainty is unavoidable. A piece of information is in general not known to be true or false, but must be annotated by shades of certainty. But what exactly is the structure of these "shades of certainty"? Are there ontologically different types of uncertainty, and, after all, how to assess, process and communicate uncertainty? One early distinction of types of uncertainty was introduced by Frank Knight in his seminal book "Risk, Uncertainty, and Profit" [Kni21] on page 19:

"Uncertainty must be taken in a sense radically distinct from the familiar notion of risk, from which it has never been properly separated.... The essential fact is that 'risk' means in some cases a quantity susceptible of measurement, while at other times it is something distinctly not of this character; and there are far-reaching and crucial differences in the bearings of the phenomena depending on which of the two is really present and operating.... It will appear that a measurable uncertainty, or 'risk' proper, as we shall use the term, is so far different from an unmeasurable one that it is not in effect an uncertainty at all."

In today's language one would describe "risk" as the uncertainty about the occurrence of events *within* a fully specified stochastic model. The "Knightian Uncertainty" is the uncertainty with regard to the correct model, what is today sometimes called model risk, especially in financial mathematics.

In the next paragraph we introduce an ontology of uncertainty, and, even more general, an ontology of indefiniteness, accompanied by a suitable terminology.

2.2 Indefiniteness

The advance of research in artificial intelligence, knowledge representation and expert systems has led to a plethora of new approaches to represent and process information (see section 2.4 for examples). This has led to confusion about the exact differences and commonalities between the different calculi, and where they are competing approaches and where they are complementary. One striking example is fuzzy logic, which is still regarded as an alternative calculus for processing uncertain information, where in fact it is a generalization of the notion of an event. This is clearly stated by Judea Pearl in [Pea00]: "Fuzzyness is orthogonal to probability theory - it focuses on the ambiguities in describing events, rather than the uncertainty about the occurrence or non-occurrence of events." Classical events are called crisp, in order to express that they are definitely defined: in a specific

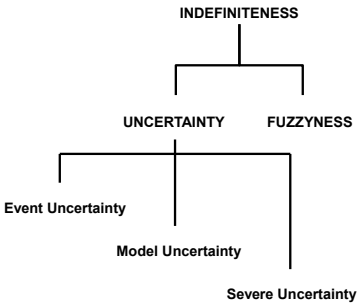

Figure 1: Ontology of Indefiniteness

situation the event has occurred or not – there are no "degrees of occurrence". The standard approach to represent a set of crisp events is a Boolean algebra. In this sense, one can say that a crisp event is an element of a Boolean algebra.

We suggest the notion "indefiniteness" for describing all sorts of non-certain, non-crisp information. This leads to the following ontology of indefiniteness:

2.3 Types of Uncertainty

Here we propose three types of uncertainty, extending the Knightian ontology:

1. Event Uncertainty (quantitatively known unknowns)
2. Model Uncertainty (qualitatively known unknowns)
3. Severe Uncertainty (unknown unknowns)

We want to illustrate these three types of uncertainty – and their principal differences – with an example taken from Bernoulli processes:

2.3.1 Event Uncertainty:

Consider the coin model with $p = \frac{1}{2}$. The question what will be the next outcome of an observation can be answered by a definite probability. In this case the probability is $\frac{1}{2}$, meaning that we are maximally unsure what will happen next, even

under a specific, complete stochastic model, but other questions can be answered with more certainty by the coin model. For example, the probability that we will observe 450 to 550 heads out of 1000 tosses of the coin is greater than 0.998. So, for this specific question the coin model delivers an answer with near certainty.

2.3.2 Model Uncertainty:

Here we assume that the observations are generated by a Bernoulli process, but with unknown success probability p. Without introducing a prior distribution for the model parameter, this implies that we only can infer probability intervals for events, for example the probability that we will observe between 45 and 55 successes out of 100 experiments is in the interval $[0, .71]$, regardless of the value of p.

2.3.3 Severe Uncertainty:

This is the "black swan" case, the possibility, that the true model is not even approximately in the set of considered models. An example would be that the true process is a deterministic switch between successes and failures, leading to a probability of 1 for the above example.

The case of severe uncertainty leads to the question of how to describe all possible models. If one requires that a model has to be an algorithmic object, the answer to this question is the set of all programs, also called program space. R. Solomonoff pioneered learning in program space in the 1960s, employing a Bayesian framework for describing model uncertainty and a prior distribution on programs inspired by Occam's razor [Sol64a, Sol64b]. Unfortunately, despite the fact that all models have to be represented by programs, the learning process devised by Solomonoff for the whole program space is not computable. The question of how to essentially retain the generality of Solomonoff's approach, but render the learning process computable has spawned a research area of its own, which is today called universal induction or algorithmic probability [Hut05, Sch09]. In section 9 we sketch how to use a combined search in program and proof space in order to get naturally defined effective instances of Solomonoff induction.

2.4 Existing Approaches to Uncertainty

Current approaches to formalize uncertainty can be characterized by a priori defining the structure of uncertainty values, and then introducing axioms for measures

on proposition algebras which should be satisfied for all valid uncertainty measures discussed in this specific approach. Additionally, in most cases at first unconditionalized measures are introduced, and only then the problem of conditionalization of such measures is discussed, often resulting in longstanding problems to construct the "right" conditionalization rule. For example, [SF02] analyzes seven different candidates for conditionalization rules in Dempster-Shafer theory, identifying pathological examples in each case. In contrast, we derive and not define the structure of uncertainty and directly axiomatize conditional uncertainty measures, thus tackling the problem of representing and processing uncertainty in one integrated approach.

The best known example certainly is probability theory (used as an inference theory in a Bayesian context [Ber85, BS94]):

- First, the domain of probability values is defined: the $[0, 1]$-interval of the real numbers.
- Second, the Kolmogorov axioms define the properties of unconditionalized probability measures.
- Third, conditionalization is defined by reduction to unconditional probability measures.

This approach has found many variants and generalizations in the literature on representing and processing uncertainty, e.g., Dempster-Shafer theory [Dem67, Sha76], Possibility theory [DP88, Dub06], Revision theory [Gär92], Ranking theory [Spo99, Spo09] or non-monotonic logic [Gin87], all exhibiting their own set of advantages and disadvantages. A survey and discussion of many of the existing approaches is given in [HSP09]. Relationships of the uncertainty calculus introduced in this thesis and existing approaches, like DS-theory or non-monotonic logic, are discussed in section 7.

3 Formalizing Uncertainty

First we have to discuss a subtle issue of terminology. Above we have used the notion "uncertainty values" to denote generalized truth values. Unfortunately, there is the following problem when using this term in a formalized context: no uncertainty about a proposition can be identified with sure knowledge, but maximal uncertainty about a proposition is *not* certainty with regard to the negation of the

proposition. The domains of truth values we want to axiomatize contain a greatest and a least element, where the greatest element should represent certainty and the least element impossibility, i.e. certainty of the negated proposition. For this reason, we adopt the notion "confidence measure" instead of uncertainty measure in the following definitions and axioms.

3.1 The Algebra of Truth Bearers

Before delving into the structure of uncertainty, we have to define the objects and their relations which are capable to take on truth values, the *truth bearers*. In a context of crisp events, i.e., after the fact it is unambiguously decidable if the event has occurred or not, the algebra of truth bearers is normally considered to be a Boolean algebra, but when truth bearers are not crisp, then another proposition algebra has to be considered, i.e., a fuzzy logic where the law of complementation is not valid: $x \vee \neg x \neq 1$, or quantum logic. The propositional algebra in quantum logic is "formally indistinguishable from the calculus of linear subspaces of a Hilbert space with respect to set products, linear sums and orthogonal complements" corresponding to the roles of and, or and not in a Boolean algebra. These linear subspaces form orthomodular lattices which in general do not satisfy the distributivity laws, see [PR08], page 128ff.

A generalization containing both fuzzy and quantum logic would be bounded lattices having an antitone involution. There seems to be no established term for this class of lattices, but they were already the topic of investigations, see for example [Cha03], page 578. We propose to call them *proto-complemented lattices*.

It would be an interesting question whether proto-complemented lattices still admit structural implications like the ring theorem (see section 4 on structure theorems for uncertainty) as is the case for Boolean algebras, or if for these generalized proposition algebras a generalization of the structure of uncertainty is necessary, too. However, in this thesis we focus on Boolean algebras as the structure of propositions. The investigation of uncertainty measures for non-Boolean proposition algebras is open to future research.

3.2 Uncertainty: the Boolean Case

A *conditional confidence measure* for a Boolean Algebra $\mathbf{U}$ and a domain of confidence values $\mathcal{C}$ is a mapping $\Gamma : \mathbf{U} \times \mathbf{U} \setminus \{\bot\} \to \mathcal{C}$. Let $A, B \in \mathbf{U}$, then the expression $\Gamma(A|B)$ reads: "the confidence value of A given B (wrt. Γ)". The

domain of confidence values is partially ordered and has a greatest ($\top\!\!\top$) and a least ($\bot\!\!\bot$) element. A *confidence space* is a triple $(\mathbf{U}, \Gamma, \mathcal{C})$. One of the following axioms (Extensibility) for confidence measures deals with relations between confidence spaces defined over different Boolean algebras. Thus it is necessary to introduce a *set of confidence spaces* all sharing the same domain of confidence values. Such a set of confidence spaces we will call a *confidence universe*, and the following axiom system is concerned with such confidence universes, and not single confidence spaces. This seemingly technical shift in perspective is essential for the formalization of natural properties like extensibility, which plays a crucial role as an intuitive axiom complementing Cox's assumptions (see section 6).

We now state seven axioms, which can be grouped in three "connective axioms" and four "infrastructure axioms", where the connective axioms concern properties of the logical connectives and the infrastructure axioms deal with basic properties of the order relations, the combinability of confidence spaces and a closure property.

3.3 Axioms for Uncertainty Measures

In the following, we use $\Gamma(A)$ as an abbreviation for $\Gamma(A|\top)$.

(Not) For all $(\mathbf{U_1}, \Gamma_1, \mathcal{C})$ and $(\mathbf{U_2}, \Gamma_2, \mathcal{C})$:
If $\Gamma_1(A_1) = \Gamma_2(A_2)$, then $\Gamma_1(\bar{A}_1) = \Gamma_2(\bar{A}_2)$.

The axiom **Not** expresses that the information in the confidence value of a statement A is sufficient to determine the confidence value of $\bar{A}$. This is justified by the requirement that every piece of information which is relevant for the confidence value of A is relevant for the confidence value of $\bar{A}$ and vice versa.

($\mathbf{And}_1$) For all $(\mathbf{U_1}, \Gamma_1, \mathcal{C})$ and $(\mathbf{U_2}, \Gamma_2, \mathcal{C})$:
If $\Gamma_1(A_1|B_1) = \Gamma_2(A_2|B_2)$ and $\Gamma_1(B_1) = \Gamma_2(B_2)$, then $\Gamma_1(A_1B_1) = \Gamma_2(A_2B_2)$.

The axiom $\mathbf{And}_1$ states that the information in the confidence values of the partial propositions determine the confidence value of the conjunction. Otherwise the confidence value of the conjunction would contain information which is not reflected in the partial propositions, although this information would be clearly relevant for at least one of them.

($\mathbf{And}_2$) For all $(\mathbf{U_1}, \Gamma_1, \mathcal{C})$ and $(\mathbf{U_2}, \Gamma_2, \mathcal{C})$:

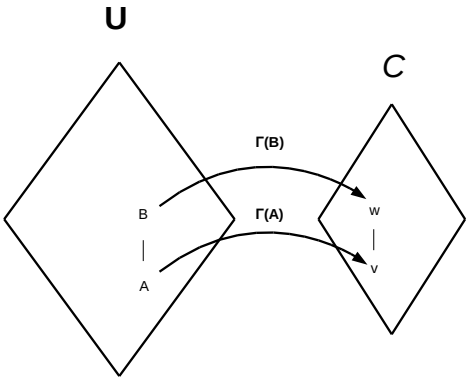

Figure 2: Ordered confidence values v and w with corresponding propositions in a suitably chosen confidence space $(\mathbf{U}, \Gamma, \mathcal{C})$.

If $\Gamma_1(A_1B_1) = \Gamma_2(A_2B_2)$ and $\Gamma_1(B_1) = \Gamma_2(B_2) \neq \perp\!\!\!\perp$, then $\Gamma_1(A_1|B_1) = \Gamma_2(A_2|B_2)$.

The axiom $\mathbf{And}_2$ ensures that all the information contained in a conditional confidence value $\Gamma(A|B)$ will be preserved in the confidence value of the conjunction $\Gamma(AB)$ when combined with the confidence $\Gamma(B)$ (unless $\Gamma(B) = \perp\!\!\!\perp$, in which case the value of $\Gamma(A|B)$ is irrelevant). Otherwise relevant information about the partial propositions would not be contained in the confidence value of the conjunction.

(Order$_1$) For all $(\mathbf{U}, \Gamma, \mathcal{C})$ and all $A, B \in \mathbf{U}$: If $A \leq B$, then $\Gamma(A) \leq \Gamma(B)$.

(Order$_2$) For all confidence values $v, w \in \mathcal{C}$ with $v \leq w$ there is a confidence space $(\mathbf{U}, \Gamma, \mathcal{C})$ with $A, B \in \mathbf{U}$ and $A \leq B$, $\Gamma(A) = v$, $\Gamma(B) = w$.

These two axioms connect the natural ordering of the Boolean algebra ($A \leq B$ iff $A \wedge B = A$) with the ordering on the confidence domain, where $\mathbf{Order}_1$ specifies the forward direction and $\mathbf{Order}_2$ specifies the backward direction (figure 2).

(Extensibility) For all $(\mathbf{U_1}, \Gamma_1, \mathcal{C})$ and $(\mathbf{U_2}, \Gamma_2, \mathcal{C})$ there is a confidence space $(\mathbf{U_3}, \Gamma_3, \mathcal{C})$, so that $\mathbf{U_3} \cong \mathbf{U_1} \otimes \mathbf{U_2}$, and for all $A_1, B_1 \in \mathbf{U}_1$, $A_2, B_2 \in \mathbf{U}_2$:

$$\Gamma_3(A_1 \otimes \top_2 \,|\, B_1 \otimes B_2) = \Gamma_1(A_1|B_1) \quad \text{and} \quad \Gamma_3(\top_1 \otimes A_2 \,|\, B_1 \otimes B_2) = \Gamma_2(A_2|B_2).$$

This axiom requires the extensibility of domains of discourse, i.e., two independently defined confidence spaces shall be embeddable into one frame of reference.

(Background) For all $(\mathbf{U}, \Gamma_1, \mathcal{C})$ and all $C \in \mathbf{U}$ there is a confidence space $(\mathbf{U}, \Gamma_2, \mathcal{C})$, so that for all $A, B \in \mathbf{U}$:

$$\Gamma_1(A|BC) = \Gamma_2(A|B)\,.$$

This closedness under conditioning assures that for every conditional confidence measure Γ_1 which is specialized by conditioning on some "background knowledge" C, there is a conditional confidence measure Γ_2 yielding the same valuations without explicitly mentioning C.

For the justification of the axioms it is important to interpret the expression $\Gamma(A|B)$ as: "*all* that can be said about the confidence of A given B (wrt. Γ)." Given this interpretation, the common justification of the connective axioms is that a violation of these axioms will necessarily lead to a loss of relevant information. Note that the axioms use only equations and inequalities between confidence values, because there are no algebraic operations defined on the domain of confidence values yet.

In order to designate this and similar axiom systems, we propose a nomenclature based on the connnective axioms. Extensionality of negation, conjunction, and disjunction is denoted as axiom N, C_1, and D_1, respectively. The reconstructibility of the confidence value of an argument of a conjunction or a disjunction, given the compositional confidence value and the confidence value of the other argument, is denoted as axiom C_2 and D_2, respectively. Using this terminology, the above introduced axiom system can be referenced as NC_{12}.

4 Structure Theorems for Uncertainty

Here we will investigate implications of the axiom system NC_{12} for the algebraic and order-theoretic structure of domains of confidence values. But before we can state and prove the two main results, the ring theorem and the total order theorem, we have to do some preparatory work by showing basic properties of confidence measures and algebraic constructions from group and ring theory.

4.1 Basic Properties of Confidence Measures

In the following we will state and prove basic properties of confidence measures of general interest, which could be useful in the context of other axiomatizations, too. Generally, the following lemmas are proved by transfering properties of Boolean algebras to the confidence domain, but such a transfer is not always possible, and the subtle parts of the proofs consist of establishing and proving the conditions enabling a transfer of properties from Boolean algebras.

First we prove a lemma stating that for every pair of confidence values there is a confidence measure and two independent events so that the confidence measure assigns the given confidence values to the independent events.

Lemma (independence lemma) For all $v, w \in \mathcal{C}$ there is a confidence space $(\mathbf{U}, \Gamma)$ with two independent events $A, B \in \mathbf{U}$, so that:

$$\Gamma(A|B) = \Gamma(A) = v, \qquad \Gamma(B|A) = \Gamma(B) = w$$

Proof: According to $\mathbf{Order}_2$, there are confidence spaces $(\mathbf{U_1}, \Gamma_1)$, $(\mathbf{U_2}, \Gamma_2)$ and events $A \in \mathbf{U_1}$ and $B \in \mathbf{U_2}$ with $\Gamma_1(A) = v$ and $\Gamma_2(B) = w$. Then axiom **Extensibility** guarantees the existence of a confidence space $(\mathbf{U_1} \otimes \mathbf{U_2}, \Gamma_3)$ with:

$$\Gamma_3(A|B) = \Gamma_3(A) = \Gamma_1(A) = v$$

and

$$\Gamma_3(B|A) = \Gamma_3(B) = \Gamma_2(B) = w.$$

■

For the definition of functions S and F see section 4.3 on proper confidence structures. In the following lemmas, the variables are assumed to take on all values in $\mathcal{C}$, i.e., these variables are implicitly universally quantified.

1. $S(S(x)) = x$
2. $x < y \Rightarrow S(x) > S(y)$
3. $F(x, y) \leq x, F(x, y) \leq y$

4. $S(F(x,y)) \geq S(x) \geq F(S(x),y)$

5. $S(\perp\!\!\!\perp) = \top\!\!\!\top, S(\top\!\!\!\top) = \perp\!\!\!\perp$

6. $\Gamma(\top|\cdot) = \top\!\!\!\top, \Gamma(\perp|\cdot) = \perp\!\!\!\perp$

Proofs:

1. Let $(\mathbf{U}, \Gamma, \mathcal{C})$ be a confidence space with $\Gamma(A) = x$ (exists according to Order_2). Then $y := \Gamma(\neg A) = S(\Gamma(A)) = S(x)$ and $S(y) = \Gamma(\neg(\neg A)) = \Gamma(A) = x$. Thus $S(y) = S(S(x)) = x$.

2. First we prove the weaker statement $x \leq y \Rightarrow S(x) \geq S(y)$. Let $(\mathbf{U}, \Gamma, \mathcal{C})$ be a confidence space with $\Gamma(A) = x, \Gamma(B) = y$, and $A \leq B$. Then $\neg B \leq \neg A$ (property of Boolean algebras). Order_1 then implies: $\Gamma(\neg B) \leq \Gamma(\neg A)$, and by axiom Not we have $S(\Gamma(B)) \leq S(\Gamma(A))$ and thus $S(y) \leq S(x)$.

The strictness can be proved by invoking lemma 1. Assume there are x, y with $x < y$ and $S(x) = S(y)$. Then S would not be an injective function, contradicting lemma 1, which states that S is an invertible function.

3. Let $(\mathbf{U}, \Gamma, \mathcal{C})$ be a confidence space with $\Gamma(A) = x, \Gamma(B) = y$, $\Gamma(A|B) = \Gamma(A)$ and $\Gamma(B|A) = \Gamma(B)$ (exists according to independence). Then $F(x,y) = F(\Gamma(A), \Gamma(B)) = F(\Gamma(A|B), \Gamma(B)) = \Gamma(AB)$. Now $AB \leq A$ and $AB \leq B$, thus applying Order_1 yields the result.

4. This can be derived by combining lemma 1 and lemma 3.

5. S is a surjective function. This can be seen by the following argument: Let x be an arbitrary confidence value, then there is a confidence space $(\mathbf{U}, \Gamma, \mathcal{C})$ with $\Gamma(A) = x$. Then $S(\Gamma(\neg A)) = \Gamma(\neg(\neg A)) = \Gamma(A) = x$. Together with lemma 1 this yields that S is a bijective function. Hence, S is a bijective and antitone (lemma 2) function, and such a function satisfies the stated property.

6. Assume there is $\Gamma(\perp|A) > \perp\!\!\!\perp$. There is a Γ' with $\Gamma'(\perp') = \perp\!\!\!\perp$ (by Order_1 and Order_2). Applying Extensibility yields: $\Gamma''(\perp \otimes \top' | A \otimes \top') = \Gamma(\perp|A) > \perp\!\!\!\perp$ and $\Gamma''(\top \otimes \perp' | A \otimes \top') = \Gamma'(\perp') = \perp\!\!\!\perp$. But $\perp \otimes \top' = \top \otimes \perp' = \perp''$. Thus we have a contradiction, proving the lemma.

4.2 Extensions of Partially Ordered Structures

Here we establish two extendability results for partially ordered structures, which hold in general and not only in the context of NC_{12}.

Group Theorem: Let $(M, +, 0, \leq)$ be a partially ordered, cancellative, commutative monoid. Then M can be extended into a partially ordered, commutative group.

Proof: This will be done by a classical algebraic construction, much like the construction of $\mathbf{Z}$ from $\mathbf{N}$. Define $G = M \times M/\sim$, where $\sim$ is an equivalence relation defined by:

$$(a, b) \sim (c, d) \Leftrightarrow a + d = c + b$$

Reflexivity and symmetry of this relation follow straightforward from the definition, but for transitivity we need associativity, commutativity, and the cancellation property of $+$. Assume the relations:

$$(a, b) \sim (c, d) \quad \text{and} \quad (c, d) \sim (e, f)$$

According to the definition, this implies the equations:

$$a + d = c + b \quad \text{and} \quad c + f = e + d$$

These equations can be transformed to:

$$(a + d) + f = (c + b) + f \quad \text{and} \quad (c + f) + b = (e + d) + b$$

Now, using associativity and commutativity of $+$, it follows that

$$(c + b) + f = (c + f) + b$$

and therefore

$$(a + d) + f = (e + d) + b$$

Again, using associativity and commutativity, the last equation can be rewritten into:

$$(a + f) + d = (e + b) + d$$

Using the cancellation property to get rid of d yields:

$$a + f = e + b$$

which is equivalent to $(a, b) \sim (e, f)$, thus establishing transitivity of $\sim$.

Next we define an operation on G, denoted by '$+_G$', which will be the extension of $+$ to a group:

$$(a, b) +_G (c, d) = (a + c, b + d)$$

We now show that $\sim$ is a *congruence relation*, i.e., compatible with the algebraic operation '$+_G$':

$$(a, b) +_G (c, d) = (a', b') +_G (c', d'), \text{ if } (\mathrm{a}, \mathrm{b}) \sim (\mathrm{a}', \mathrm{b}') \text{ and } (\mathrm{c}, \mathrm{d}) \sim (\mathrm{c}', \mathrm{d}')$$

The operation '$+_G$' is commutative, so it suffices to analyze the case where $c = c'$ and $d = d'$.

$$(a, b) \sim (a', b') \Rightarrow a + b' = a' + b \Rightarrow (a + b') + c = (a' + b) + c$$

$$\Rightarrow ((a + b') + c) + d = ((a' + b) + c) + d$$

$$\Rightarrow (a + c) + (b' + d) = (a' + c) + (b + d)$$

The last equation is derived by using the associativity and commutativity of $+$. This equation is by definition equivalent to:

$$(a + c, b + d) \sim (a' + c, b' + d)$$

which in turn implies according to the definition of $+_N$:

$$(a, b) +_G (c, d) \sim (a', b') +_G (c, d)$$

The last equation states that the replacement of an argument of '$+_G$' (the extension of $+$) by an equivalent one yields an equivalent result, i.e., '$\sim$' is a congruence relation with regard to '$+_G$'.

After we have established that $+_G$ is a well-defined operation on G, we have to prove that it has the group properties, i.e., associativity, existence of neutral element, and the existence of inverse elements. Associativity of $+_G$ directly reduces to the associativity of $+$:

$$((a,b) +_G (c,d)) +_G (e,f) = (a+c+e, b+d+f) = (a,b) +_G ((c,d) +_G (e,f))$$

and $(0,0)$ is a neutral element. Finally, (b,a) is an inverse of (a,b):

$$(a,b) +_G (b,a) = (a+b, b+a) \sim (0,0)$$

Commutativity of $+_G$ follows directly from the commutativity of $+$, much like associativity. We have now established that $(G, +_G, [0,0])$ forms a commutative group. This closes the algebraic part of the group theorem. It remains to extend the partial order of the monoid and prove the compatibility with the group operation $+G$. We extend the partial order $\leq$ of M to a partial order $\leq_G$ on G by the following definition:

$$[a,b] \leq_G [c,d] \quad \Leftrightarrow \quad a+d \leq c+b$$

That this is a well-defined relation can be seen in the same way as we have shown the well-definedness of $+_G$: just replace $=$ by $\leq$ in the derivation, and it follows that the relation $\leq_G$ does not depend on the specific element $[a,b]$ representing the equivalence class. Now that we know that $\leq_G$ is a well-defined relation on G, we have to prove that it is a partial order, i.e., reflexive, transitive, and anti-symmetric. Reflexivity and transitivity are analoguously shown to the reflexivity and transitivity of $\sim$. It remains to establish the anti-symmetry:

$$[a,b] \leq_G [c,d] \text{ and } [\mathrm{c,d}] \leq_{\mathrm{G}} [\mathrm{a,b}] \;\Rightarrow\; [\mathrm{a,b}] = [\mathrm{c,d}]$$

This can be seen by translating the propositions about G back to propositions about M:

$$[a,b] \leq_G [c,d] \quad \Rightarrow \quad a+d \leq c+b, \quad [c,d] \leq_G [a,b] \quad \Rightarrow \quad c+b \leq a+d$$

Because $\leq$ is a partial order on M it is an anti-symmetric relation, leading to:

$$a+d = c+b,$$

which is equivalent to $[a,b] \sim [c,d]$. Hence $\leq_G$ is anti-symmetric, too.

Finally, we have to show that the partial order $\leq_G$ is compatible with the group operation $+_G$:

$$[a,b] \leq_G [c,d] \quad \Rightarrow \quad [a,b] + [e,f] \leq_G [c,d] + [e,f]$$

We prove this by translating the left-hand side into a proposition about M, then transforming this proposition according to the properties of M, and finally translate back to right-hand side proposition about G:

$$a+d \leq c+b \quad \Rightarrow \quad a+d+e+f \leq c+b+e+f \quad \Rightarrow \quad (a+e)+(d+f) \leq (c+e)+(b+f)$$

$$\Rightarrow \quad [a+e, b+f] \leq_G [c+e, d+f] \quad \Rightarrow \quad [a,b] + [e,f] \leq_G [c,d] + [e,f]$$

This establishes the compatibility of the partial order $\leq_G$ with the group operation $+_G$, and thus finishes the proof of the group theorem.

■

The next extension result concerns the extension of semi-rings, i.e. rings where the addition operation is only a monoid, into a ring, i.e., where the addition operation is a group.

Semi-Ring Theorem: Let $(R, +, \cdot, 0, 1, \leq)$ be a partially ordered, sum-cancellative, commutative semi-ring, which satisfies the following compatibility condition: if $a \geq b$, then there is a c with $a = b + c$. Then R can be extended into a partially ordered, commutative ring.

Proof: The algebraic part of this theorem is a well-known construction of a ring from a semiring, see, for example, theorem 8.7 in [HW96]. It remains to extend

the partial order and prove its compatibility with the original partial order of the semiring. The extended partial order is defined as follows:

$$[a,b] \leq_R [c,d] \quad \Leftrightarrow \quad a+d \leq c+b$$

The compatibility with addition can be established analoguously to the proof of the compatibility of the extended partial order in the group theorem, i.e.:

$$a+d \leq c+b \;\; \Rightarrow \;\; a+d+e+f \leq c+b+e+f \;\; \Rightarrow \;\; (a+e)+(d+f) \leq (c+e)+(b+f)$$

$$\Rightarrow \quad [a+e, b+f] \leq_R [c+e, d+f] \quad \Rightarrow \quad [a,b]+[e,f] \leq_R [c,d]+[e,f]$$

It remains to show the compatibility of the partial order with multiplication, i.e.:

$$[a,b] \leq_R [c,d] \;\text{ and }\; [\mathrm{e,f}] \geq_\mathrm{R} [0,0] \quad \Rightarrow \quad [\mathrm{e,f}] \cdot [\mathrm{a,b}] \leq_\mathrm{R} [\mathrm{e,f}] \cdot [\mathrm{c,d}]$$

Whenever $[e,f] \geq_R [0,0]$, there is a $e' \geq 0$ with $[e,f] \sim [e',0]$. This is ensured by the compatibility condition. Because $\sim$ is a congruence relation with regard to multiplication, we get:

$$[e,f] \cdot [a,b] = [e',0] \cdot [a,b]$$

and

$$[e,f] \cdot [c,d] = [e',0] \cdot [c,d]$$

Now starting from $[a,b] \leq_R [c,d]$, we get:

$$[a,b] \leq_R [c,d] \quad \Rightarrow \quad a+d \leq c+b \quad \Rightarrow e' \cdot (a+d) \leq e' \cdot (c+b)$$

$$\Rightarrow \quad e' \cdot a + e' \cdot d \leq e' \cdot c + e' \cdot b \quad \Rightarrow \quad [e' \cdot a, e' \cdot b] \leq_R [e' \cdot c, e' \cdot d]$$

$$\Rightarrow \quad [e',0] \cdot [a,b] \leq_R [e',0] \cdot [c,d]$$

Thus we have compatibility of the extended partial order on the ring with the extended multiplication, which completes the proof of the semiring theorem.

■

4.3 Proper Confidence Structures

The proof of the ring theorem will be divided into two parts. First we introduce an intermediate structure which we will call *proper confidence structure*, or PCS for short. We then show that every model of NC_{12} is a PCS. A PCS is a natural interface between lower and higher structural properties of confidence measures, which can be useful in future investigations of alternative axiom systems.

Definition: A **proper confidence space** is a seven-tuple $(\mathcal{C}, F, G, S, \leq, \perp\!\!\!\perp, \top\!\!\!\top)$, where $\mathcal{C}$ is a partially ordered set with smallest value $\perp\!\!\!\perp$ and largest value $\top\!\!\!\top$, $F : \mathcal{C} \times \mathcal{C} \to \mathcal{C}$, $S : \mathcal{C} \to \mathcal{C}$, $\mathcal{E} = \{(x, y) \in \mathcal{C} \times \mathcal{C} | x \leq S(y)\}$ and $G : \mathcal{E} \to \mathcal{C}$. Moreover, F and G are symmetric and associative, F distributes over G, F and G are increasing in their arguments and S is decreasing. Additionally, $G(x, \perp\!\!\!\perp) = x$, $F(\perp\!\!\!\perp, x) = \perp\!\!\!\perp$, $F(x, \top\!\!\!\top) = x$, $S(S(x)) = x$, and $S(\perp\!\!\!\perp) = \top\!\!\!\top$.

Theorem: All models of NC_{12} are proper confidence structures.

Proof: The proof consists of the following steps:

1. Existence of Functions F and S
2. F is associative, commutative and cancellative.
3. F can be extended to a group
4. Definition of Function G via F, S, representing disjunction.
5. Well-definedness of G.
6. F is distributive over G.

The first step is to show that axioms **Not** and $\mathbf{And}_1$ imply the existence of functions $S : \mathcal{C} \to \mathcal{C}$ and $F : \mathcal{C} \times \mathcal{C} \to \mathcal{C}$, where S relates the confidence value of a proposition to the confidence value of its negation, and F transforms the confidence value of two propositions to the confidence value of their conjunction.

Lemma There is a function S on the set $\mathcal{C}$ of confidence values, so that for all confidence spaces $(\mathbf{U}, \Gamma)$ and all $A, B \in \mathbf{U}$:

$$\Gamma(\bar{A}|B) = S(\Gamma(A|B))$$

Proof: Let $v \in \mathcal{C}$ be a confidence value, so that there is a confidence space $(\mathbf{U_1}, \Gamma_1)$ and $A_1 \in \mathbf{U_1}$ with $\Gamma_1(A_1) = v$. Then define:

$$S(v) = \Gamma_1(\bar{A}_1)$$

This function is well-defined, because whenever there is another confidence space $(\mathbf{U_2}, \Gamma_2)$ having v as value of the confidence measure Γ_2, say $\Gamma_2(A_2|B_2) = v$, then axiom **Not** assures that $\Gamma_2(\bar{A}_2|B_2) = \Gamma_1(\bar{A}_1)$. That is, the value of S does not depend on the specific choice of confidence space having v as a value. Additionally, S is a total function by axiom $\mathbf{Order}_2$, which enforces that for every $v \in \mathcal{C}$ there is at least one confidence space taking v as a value.

■

The analog will be proved for conjunction by introducing a binary function F on $\mathcal{C}$. Note that a proposition B and a *conditional* propostion $A|B$ are related by F.

Lemma There is a binary function F on the set $\mathcal{C}$ of confidence values, so that for all confidence spaces $(\mathbf{U}, \Gamma)$ and all $A, B, C \in \mathbf{U}$:

$$\Gamma(AB|C) = F(\Gamma(A|BC), \Gamma(B|C))$$

Proof: According to the independence lemma, there is for all $v, w \in \mathcal{C}$ a confidence space $(\mathbf{U_1}, \Gamma_1)$ with $\Gamma_1(A_1|B_1) = \Gamma_1(A_1) = v$ and $\Gamma_1(B_1) = w$. Define F as follows:

$$F(v, w) = \Gamma_1(A_1 B_1)$$

The well-definedness is implied by axiom $\mathbf{And}_1$, the value of $F(v, w)$ does not depend on the confidence space and events having v and w as confidence values. The totality of F is given by the independence lemma, which works for all pairs $v, w \in \mathcal{C}$.

■

Lemma The function F is associative.

Proof: Let $x, y, z \in \mathcal{C}$ and $(\mathbf{U}, \Gamma)$ a confidence space with $A, B, C \in \mathbf{U}$ and $\Gamma(A|BC) = \Gamma(A) = x$, $\Gamma(B|C) = \Gamma(B) = y$ and $\Gamma(C) = z$. Such a confidence measure always exists according to the independence lemma. Then it follows:

$F(F(x, y), z) = F(F(\Gamma(A), \Gamma(B)), \Gamma(C)) = F(F(\Gamma(A|BC), \Gamma(B|C)), \Gamma(C)) =$

$F(\Gamma(AB|C), \Gamma(C)) = \Gamma(ABC) = F(\Gamma(A|BC), \Gamma(BC)) =$

$$F(\Gamma(A), F(\Gamma(B|C), \Gamma(C))) = F(\Gamma(A), F(\Gamma(B), \Gamma(C))) = F(x, F(y, z))$$

■

Lemma The function F is commutative.

Proof: Let $x, y \in \mathcal{C}$ and $(\mathbf{U}, \Gamma)$ a confidence space with $A, B \in \mathbf{U}$ and $\Gamma(A|B) = \Gamma(A) = x$ and $\Gamma(B) = y$. Again the independence lemma guarantees the existence of such a confidence measure. Then it follows:

$F(x, y) = F(\Gamma(A|B), \Gamma(B)) = \Gamma(AB) = \Gamma(BA) =$

$F(\Gamma(B|A), \Gamma(A)) = F(\Gamma(B), \Gamma(A)) = F(x, y).$

■

Lemma The function F is cancellative.

Proof: Let $x, y, a \in \mathcal{C}$ and $(\mathbf{U}, \Gamma)$ a confidence space with $A, B, C, D \in \mathbf{U}$ and $\Gamma(A|C) = \Gamma(A) = x$, $\Gamma(B|D) = \Gamma(B) = y$, and $\Gamma(C) = \Gamma(D) = a$, again using the independence lemma. Now:

$F(x, a) = \Gamma(AC)$ and $F(y, a) = \Gamma(BD)$.

Thus, $F(x, a) = F(y, a)$ implies $\Gamma(AC) = \Gamma(BD)$. Invoking And_2, we get $\Gamma(A|C) = \Gamma(B|D)$, i.e. $x = y$.

■

We now have established that $(\mathcal{C} \backslash \{\perp\!\!\!\perp\}, F, \top\!\!\!\top, \leq)$ is a partially ordered, cancellative, commutative monoid, so we can invoke the group theorem to extend $\mathcal{C}$ into a partially ordered group. The analogy to the classical structures is the extension of the $(0, 1]$-interval of $\mathbf{Q}$ with multiplication as the single operation to all positive rational numbers. We now introduce a new operation which will turn out to be the analogue of addition on $\mathbf{Q}$.

We now define a partial function G on $\{(x, y) | x, y \in \mathcal{C}, x \leq S(y)$, which later can be seen to be connected to disjunctions of propositions. Using S, F, and F^{-1}, the function G is defined as follows:

$$G(x, y) = S(F(S(F(x, F^{-1}(S(y)))), S(y)))$$

In order to illustrate this definition, we note that G essentially has to solve the problem to represent addition with the functions $x * y$, $1 - x$, and $1/x$. Using these functions, G becomes:

$$1 - (1 - \tfrac{x}{1-y})(1 - y)$$

which reduces to addition.

First we have to show that this is a well-defined function. For this, we must show that on the domain of G the expression $F(x, F^{-1}(S(y)))$ is in $\mathcal{C}$, because the S-function is still only defined on $\mathcal{C}$, and not on the group extension of $\mathcal{C}$.

Lemma: $\forall x, y \in \mathcal{C} : x \leq S(y) \;\Rightarrow\; F(x, F^{-1}(S(y))) \in \mathcal{C}$.

Proof: With Order_2 and $x \leq S(y)$ it follows that there is a confidence space $(\mathbf{U}, \Gamma, \mathcal{C}$ with $A, B \in \mathbf{U}$, $A \leq B$, $\Gamma(A) = x$, and $\Gamma(B) = S(y)$. Now, because of $A \leq B$, it holds that $\Gamma(AB) = \Gamma(A) = x$. Let $\Gamma(A|B) = z$, which is uniquely determined according to And_2. Then we have:

$$x = F(z, S(y)) \Leftrightarrow z = F(x, F^{-1}(S(y)))$$

z is in the range of a confidence measure ($z = \Gamma(A|B)$), thus $z \in \mathcal{C}$, which proves the lemma.

■

Lemma: F distributes over G: $F(x, G(y, z)) = G(F(x, y), F(x, z)), \forall\ y \leq S(z)$.

Proof: Let $(\mathbf{U}, \Gamma_1, \mathcal{C})$ be a confidence space with $\Gamma_1(B) = y$, $\Gamma_1(C') = S(z)$, and $B \leq C'$ (exists according to Order_2). Let $C = \neg C'$, then $BC = \bot$. Further, let $(\mathbf{U}, \Gamma_2, \mathcal{C})$ be a confidence space with $\Gamma_2(A) = x$ and $(\mathbf{U}, \Gamma_3, \mathcal{C})$ a confidence space with $\Gamma_3(A|B) = \Gamma_3(A) = x$, $\Gamma_3(A|C) = \Gamma_3(A)$, and $\Gamma_3(A|B \vee C) = \Gamma_3(A)$ (exists according to Extensibility). Then we get the following chain of equations:

$$F(x, G(y, z)) = F(\Gamma_3(A), G(\Gamma_3(B), \Gamma_3(C))) = F(\Gamma_3(A), \Gamma_3(B \vee C))$$

$$= F(\Gamma_3(A|B \vee C), \Gamma_3(B \vee C)) = \Gamma_3(A(B \vee C))$$

$$= \Gamma_3(AB \vee AC) = G(\Gamma_3(AB), \Gamma_3(AC))$$

The last equation holds because of $BC = \bot$, i.e., $ABAC = \bot$, too.

$$G(F(\Gamma_3(A|B), \Gamma_3(B)), F(\Gamma_3(A|C), \Gamma_3(C)))$$

$$= G(F(\Gamma_3(A), \Gamma_3(B)), F(\Gamma_3(A), \Gamma_3(C)))$$

$$= G(F(x, y), F(x, z))$$

■

We now have established the algebraic properties of a proper confidence structure. It remains to show that the order properties of a PCS are satisfied, too.

Lemma: F is increasing in its arguments if the other argument is not $\perp\!\!\!\perp$.

Proof: First we show that F is non-decreasing. Let $x, y, z \in \mathcal{C}$ and $y \leq z$. Order_2 implies the existence of confidence spaces $(\mathbf{U_1}, \Gamma_1, \mathcal{C}), (\mathbf{U_2}, \Gamma_2, \mathcal{C})$ with $A, B \in \mathbf{U_1}$, $A \leq B$, $\Gamma_1(A) = y, \Gamma_1(B) = z$ and $C \in \mathbf{U}$ with $\Gamma_2(C) = x$. By combining both confidence spaces using the extensibility axiom into a confidence space $(\mathbf{U_3}, \Gamma_3, \mathcal{C})$ we can conclude as follows:

$F(x, y) = F(\Gamma_3(C), \Gamma_3(A)) = F(\Gamma_3(C|A), \Gamma_3(A)) = \Gamma(CA) \leq \Gamma(CB) =$
$= F(\Gamma_3(C|B), \Gamma_3(B)) = F(\Gamma_3(C), \Gamma_3(B)) = F(x, z)$

The step $\Gamma(CA) \leq \Gamma(CB)$ uses axiom Order_1, which is possible because $A \leq B$ implies $CA \leq CB$. The strictness now follows from axiom And_2 for all $x \neq \perp\!\!\!\perp$.

■

Lemma: G is increasing.

Proof: Let $x, y, z \in \mathcal{C}$ with $x \leq S(y)$, $x \leq S(y)$, and $y \leq z$. Again, using axioms Order_2 and Extensibility, there is a confidence space $(\mathbf{U}, \Gamma, \mathcal{C})$ with $CA = \perp, CB = \perp, A \leq B$. It follows:

$G(x, y) = G(\Gamma(C), \Gamma(A)) = \Gamma(C \vee A) \leq \Gamma(C \vee B) = G(\Gamma(C), \Gamma(B)) = G(x, z)$

Strictness then follows from the strictness of S and F.

■

Lemma: S is decreasing.

Proof: See basic property 2.

■

This completes the proof of the PCS property.

4.4 Ring Theorem

All models of the axiom system NC_{12} can be endowed with a ring structure[1], and the confidence measures satisfy the analogons of the Kolmogorov axioms wrt. the ring operations. This is stated precisely in the following theorem:

Ring Theorem: The domain of confidence values $\mathcal{C}$ of a confidence universe satisfying the axiom system NC_{12} can be embedded into a partially ordered commutative ring $(\hat{\mathcal{C}}, 0, 1, \oplus, \odot, \leq)$. Let $\hat{\cdot} : \mathcal{C} \to \hat{\mathcal{C}}$ be the embedding map, then the following holds:

$$\hat{\perp\!\!\!\perp} = 0\,, \quad \hat{\top\!\!\!\top} = 1\,, \quad \forall v, w \in \mathcal{C} :\ v \leq w \ \Leftrightarrow\ \hat{v} \leq \hat{w}\,.$$

Furthermore, all confidence measures Γ of the confidence universe satisfy:

$$\hat{\Gamma}(\top) = 1\,, \tag{1}$$

$$\hat{\Gamma}(A \vee B) = \hat{\Gamma}(A) \oplus \hat{\Gamma}(B)\,, \qquad \text{if } AB = \perp\,, \tag{2}$$

$$\hat{\Gamma}(A \wedge B) = \hat{\Gamma}(A|B) \odot \hat{\Gamma}(B)\,. \tag{3}$$

The proof of the ring theorem will be conducted along the following lines:

1. G can be extended into a total function.
2. The extended G is cancellative.
3. F, G and $\leq$ build a partially ordered, cancellative, commutative semiring.
4. Application of the the semi-ring theorem leads to a partially ordered, commutative ring.

4.4.1 Extension of G into a total function

We start by showing some basic properties of G with regard to its domain of definition (in order to increase readability we replace G by an infix $+$ and F by an infix $\cdot$):

[1] possibly containing zero divisors, see section 6.3.

Lemma R1: If $a_1 + a_2$ is defined and $a_1 \geq b_1$, $a_2 \geq b_2$, then $b_1 + b_2$ is defined.

Proof: $b_1 \leq a_1 \leq S(a_2) \leq S(b_2)$, where for the last inequality we used basic property 2.

■

Lemma R2: If $e \neq 0, 1$ and $f = e \cdot S(e)$, then $f \cdot a + f \cdot b$ is defined.

Proof: $f = e \cdot S(e) \leq S(e) \leq S(e \cdot S(e)) = S(f)$. Here we used basic properties 2 and 3 of F and S. Thus $f + f$ is defined. With basic property 3 we get $f \cdot a \leq f$ and $f \cdot b \leq f$. Hence lemma R1 is applicable, proving that $f \cdot a + f \cdot b$ is defined.

■

Lemma R3: For every sequence $(a_i)_1^n$ there is a non-zero c_n depending only on n such that $c_n \cdot a_1 + c_n \cdot a_2 + ... + c_n \cdot a_n$ is defined.

Proof: For any non-trivial confidence value e, choose $c = e \cdot S(e)$ and $c_n = c^{ld(n)}$. Use lemma R2 inductively on half sequences.

■

Currently, G is only a partial function. Lemma R3 can now be used to extend G into a total function by multiplying a pair of confidence values, which are not in the domain $\mathcal{E}$ of G by a "small" value, shifting the pair of values into the domain of G. After the application of G to the transformed pair of confidence values, the result is multiplied by the inverse of the "small" value, giving the final result. Note that in the extended structure $\mathcal{C}^1$ there are multiplicative inverses for all elements, because by construction $\mathcal{C}^1$ is a group wrt. F.

Definition: G will be extended to $\mathcal{C}^1$ by the following defintion: $(a, b) + (c, d) = (f \cdot a \cdot d + f \cdot c \cdot b, f \cdot b \cdot d)$, choosing f according to lemma R3 so that the expressions are defined.

The definition is analogue to the definition of the sum of two rational numbers, with the only difference that both numerator and denominator are multiplied by f in order to assure the well-definedness of the sum in the numerator. But the final result will not depend on the specific f, because it cancels out.

4.4.2 Cancellation property of G

We now show that the extended G has the cancellation property, which ensures that we can define additive inverses:

Lemma R4 (cancellation property of G): For all $x, y, z \in \mathcal{C}^1$ with $y \leq S(x)$ and $z \leq S(x)$ it holds:

$$y + x = z + x \quad \Rightarrow \quad y = z \qquad \text{and} \qquad x + y = x + z \quad \Rightarrow \quad y = z$$

Proof: First we show the cancellation property of G on $\mathcal{C}$ whenever both sides of the equation are defined. The cancellation property of the partial G can be reduced to the corresponding properties of S and F (assuming $x \neq \top\!\!\top$). We start with proving right cancellation:

$y + x = z + x$

$\Leftrightarrow$ (by defintion of G)

$S(F(S(F(y, F^{-1}(S(x)))), S(x))) = S(F(S(F(z, F^{-1}(S(x)))), S(x)))$

$\Leftrightarrow$ (by injectivity of S)

$F(S(F(y, F^{-1}(S(x)))), S(x)) = F(S(F(z, F^{-1}(S(x)))), S(x))$

$\Leftrightarrow$ (by cancellation property of F, $S(x) \neq \perp\!\!\!\perp$)

$S(F(y, F^{-1}(S(x)))) = S(F(z, F^{-1}(S(x))))$

$\Leftrightarrow \quad F(y, F^{-1}(S(x)) = F(z, F^{-1}(S(x))$

$\Leftrightarrow \quad y = z$

The case $x = \top\!\!\top$ enforces both y, z to be $\perp\!\!\!\perp$ by the inequalities $y \leq S(x)$ and $z \leq S(x)$, thus yielding the cancellation property in this case, too.

Left cancellations then follows from the symmetry of G.

Let us now analyse the equation $(y_1, y_2) + (x_1, x_2) = (z_1, z_2) + (x_1, x_2)$ on $\mathcal{C}^1$. According to the definition this is equivalent to

$$(f \cdot y_1 \cdot x_2 + f \cdot x_1 \cdot y_2, f \cdot y_2 \cdot x_2) = (f \cdot z_1 \cdot x_2 + f \cdot x_1 \cdot z_2, f \cdot z_2 \cdot x_2)$$

$$(f \cdot y_1 \cdot x_2 + f \cdot x_1 \cdot y_2) \cdot (f \cdot z_2 \cdot x_2)) = (f \cdot z_1 \cdot x_2 + f \cdot x_1 \cdot z_2) \cdot (f \cdot y_2 \cdot x_2)$$

$$f^2 \cdot x_2^2 \cdot y_1 \cdot z_2 + f^2 \cdot x_1 \cdot x_2 \cdot y_2 \cdot z_2 = f^2 \cdot x_2^2 \cdot y_2 \cdot z_1 + f^2 \cdot x_1 \cdot x_2 \cdot y_2 \cdot z_2$$

Using the cancellation property of '+' and '$\cdot$', we get:

$$y_1 \cdot z_2 = z_1 \cdot y_2 \ , \quad \text{which is equivalent to} \quad (y_1, y_2) = (z_1, z_2) \ .$$

■

4.4.3 Semiring property of G and F

A semiring is similar to a ring, but without the requirement that each element must have an additive inverse. We have already shown that F and G are monoids on $\mathcal{C}^1$. It remains to show that F distributes over G on $\mathcal{C}^1$.

Lemma R5: For all $x, y, z \in \mathcal{C}^1$ it holds: $(x + y) \cdot z = x \cdot z + y \cdot z$.

Proof: Let us (u_1, u_2) set equal to the right side of the equation of the lemma. Then we have:

$$(u_1, u_2) = (x_1 \cdot z_1, x_2 \cdot z_2) + (y_1 \cdot z_1, y_2 \cdot z_2)$$

$$u_1 \cdot x_2 \cdot y_2 \cdot z_2 = u_2 \cdot x_1 \cdot z_1 \cdot y_2 + u_2 \cdot y_1 \cdot z_1 \cdot x_2$$

$$u_1 \cdot x_2 \cdot y_2 \cdot z_2 = u_2 \cdot z_1 \cdot (x_1 \cdot y_2 + y_1 \cdot x_2)$$

If we set (u_1, u_2) equal to the left side of the equation of the lemma, we get:

$$(u_1 \cdot z_2, u_2 \cdot z_1) = (x_1 \cdot y_2 + y_1 \cdot x_2, x_2 \cdot y_2)$$

$$u_1 \cdot x_2 \cdot y_2 \cdot z_2 = u_2 \cdot z_1 \cdot (x_1 \cdot y_2 + y_1 \cdot x_2)$$

This establishes distributivity on $\mathcal{C}^1$.

■

4.4.4 Extension of G and F into a ring

We now can finish the proof of the ring theorem by invoking the semi-ring theorem. $(\mathcal{C}^1, +, \cdot, \perp\!\!\!\perp, \top\!\!\!\top, \leq)$ is a partially ordered, sum-cancellative, commutative semi-ring and Order_2 implies the compatibility condition. According to the semi-ring theorem we can extend such a semi-ring into a partially ordered, commutative ring. This finishes the proof of the ring theorem.

■

The ring theorem can be seen as a generalization of the Kolmogorov axioms of probability theory, as it restates these axioms, only the algebraic operations over the reals are replaced with the corresponding ring operations. Thus it follows immediately that classical probability theory is a model of the axiom system NC_{12}. But there are more general models of NC_{12}, containing uncomparable or infinitesimal elements. Hence the ring theorem characterizes not probability theory alone as the algebra of uncertainty, but a set of more general models containing probability theory as a special case. This leads to the interesting question whether there is a universally embedding confidence ring, i.e., a confidence ring containing all other confidence rings as substructures. If such a universally embedding confidence ring exists, then the [0,1]-interval of this ring would be the most general domain of uncertainty values, and could be, at least for theoretical considerations, regarded as the "default model" of the axiom system NC_{12}. In fact, in the case of totally ordered fields such a universally embedding structure exists and is called the field of surreal numbers **No**. It was introduced by J. Conway and D. Knuth in the 1970s [Con76, Knu74, Gon86]. However, the question of the existence of a universally embedding confidence ring is open to future research.

4.5 Probability theory as a model of NC_{12}

Here we discuss in detail that classical probability theory, i.e. the theory that uses the real [0,1]-interval as valuation domain and the Kolmogorov axioms in order to define the properties of probability measures, is a model of the axiom system NC_{12}:

Not

If $P_1(A) = P_2(B) = \alpha$, then by additivity we have $P_1(\bar{A}) = 1 - \alpha = P_2(\bar{B})$.

And$_1$

If for two probability measures we have $P_1(A_1|B_1) = P_2(A_2|B_2)$, i.e., these values are defined and thus $P_1(B_1) = P_2(B_2) \neq 0$, then by applying the definition of conditional probability we get $P_1(A_1B_1) = P_2(A_2B_2)$.

And$_2$

If for two probability measures we have $P_1(A_1B_1) = P_2(A_2B_2)$, and $P_1(B_1) = P_2(B_2) \neq 0$, then by applying the defintion of conditional probability we get $P_1(A_1|B_1) = P_2(A_2|B_2)$.

Order$_1$

Whenever $A \geq B$, we have that $A \setminus B \cup B = A$. Then additivity implies $P(A) = P(A \setminus B \cup B) = P(A \setminus B) + P(B) \geq P(B)$.

Order$_2$

Take as Boolean algebra the algebra with three elementary events A_1, A_2, A_3 (which is a σ-algebra, too), and define for arbitrary $\alpha, \beta \in [0, 1]$ with $\alpha \geq \beta$:

$$P(A_1) = \beta, P(A_2) = \alpha - \beta, P(A_3) = 1 - \alpha.$$

This defines a probability measure having the events $A := A_1 \cup A_2 \geq B := A_1$ and $P(A) = \alpha, P(B) = \beta$.

Extensibility

This axiom can be interpreted as the product measure theorem of probability theory: for each pair of probability spaces $(\Omega_1, \mathbf{A_1}, P_1)$ and $(\Omega_2, \mathbf{A_2}, P_2)$ there is a product space $(\Omega_1 \times \Omega_2, \mathbf{A_1} \otimes \mathbf{A_2}, P_1 \otimes P_2)$ with:

$$(P_1 \otimes P_2)(A_1 \times A_2) = P_1(A_1) \cdot P_2(A_2),$$

for all $A_1 \in \mathbf{A_1}, A_2 \in \mathbf{A_2}$. This product space satisfies the conditions of the Extensibility axiom.

Background

It is a basic property of conditional probability measures to satisfy the Kolmogorov axioms when interpreted as unconditional probability measures, i.e., the unconditional measure P_1 defined as $P_1(\cdot) = P_0(\cdot|A)$ satisfies the Kolmogorov axioms. Thus the Background axiom is valid in probability theory.

Hence all seven axioms of NC_{12} are valid in probability theory, so probability theory is a model of NC_{12}. This also implies that Bayesian inference is an admissible uncertainty calculus, it just does not use the full generality, like uncomparable or infinitesimal elements, allowed by NC_{12}.

4.6 Total Order Theorem

An interesting consequence of the ring theorem arises when we constrain the partial order on the confidence domain to have the lattice property (i.e., for each pair of elements there exist join and meet):

Total Order Theorem: If the axiom system NC_{12} is extended by the requirement that the partial order of the confidence values should have the lattice property, then the confidence values are totally ordered.

Proof: We first show some general properties of lattice-ordered commutative groups:

(LOG1) $-(x \wedge y) = (-x \vee -y)$

Let $z = x \wedge y$, i.e., $z \leq x$ and $z \leq y$. Thus $-z \geq -x$ and $-z \geq -y$. So $-z$ is an upper bound of $-x$ and $-y$, i.e., $-z \geq -x \vee -y$. Assume that there is a smaller upper bound u of $-x$ and $-y$, i.e., $-z > u \geq -x \vee -y$. Then $z < -u < x \wedge y$, in contradiction to the greatest lower bound property of z. This proves that there is no smaller upper bound for $-x$ and $-y$, i.e., $-z = -x \vee -y$.

(LOG2) $-(x \vee y) = (-x \wedge -y)$

This is the dual property and can be proven in the same way by reversing the inequalities.

(LOG3) $(x \wedge y) + z = (x + z) \wedge (y + z)$

Let $u = x \wedge y$, i.e., $u \leq x$ and $u \leq y$. It follows that $u + z \leq x + z$ and $u + z \leq y + z$. So $u + z \leq (x + z) \wedge (y + z)$. Assuming that there is a v greater than $u + z$, but still a lower bound for $x + z$ and $y + z$ would imply that $v - z$ is a greater lower bound of x and y than u. This contradiction proves the proposition.

(LOG4) $(x \vee y) + z = (x + z) \vee (y + z)$

This is the dual property and can be proven in the same way by reversing the inequalities.

(LOG5) $x + y = (x \wedge y) + (x \vee y)$

This proposition can be proved using LOG1 and LOG4:

$$x + y = x + y + (x \wedge y) - (x \wedge y) = x + y + (x \wedge y) + (-x \vee -y) = (x \wedge y) + (x \vee y)$$

The first step is adding a zero, the second step applies LOG1, and in the last step, beside applying LOG4, the commutativity of $\wedge$ is used.

Using these properties, one can show that for a lattice-ordered ring we have

$$(x - (x \wedge y) \cdot (y - (x \wedge y)) = 0$$

This can be shown by the following sequence of equations:

$$(x - (x \wedge y) \cdot (y - (x \wedge y)) = x \cdot y - x \cdot (x \wedge y) - y \cdot (x \wedge y) + (x \wedge y)^2 =$$

$$x \cdot y - (x + y) \cdot (x \wedge y) + (x \wedge y)^2 = x \cdot y - ((x \wedge y) + (x \vee y)) \cdot (x \wedge y) + (x \wedge y)^2$$

$$x \cdot y - (x \vee y) \cdot (x \wedge y) = x \cdot y - x \cdot y = 0$$

The equality of $(x \vee y) \cdot (x \wedge y)$ and $x \cdot y$ follows from LOG5 by noting that the $(0, 1]$-interval of a confidence ring is part of a subgroup of the multiplicative structure of the confidence ring, shown in the proof of the ring theorem. Hence, the product together with $\wedge$ and $\vee$ forms a lattice-ordered group, and therefore LOG5 is applicable.

Now assume that x, y are from the $(0, 1]$-interval of our confidence ring and are not comparable, i.e. $x \not\leq y$ and $y \not\leq x$. Then the factors in the above equation are non-zero, but their product is zero, i.e., they are zero-divisors. But according to And_2, there cannot be zero-divisors in the $(0, 1)$-interval. This contradiction proves that the assumption that x and y are not comparable is false, hence the order is total.

■

This is surprising insofar that lattice order in general is far from total order, but in the context of the above axioms, this weak order property has a strong implication. This result can be summarized as a trilemma: the three properties of partial order, lattice-order, and field structure of the domain of confidence values cannot be satisfied all at the same time. In future investigations one has to decide which of these three properties is most dispensable.

5 Uncertainty Decomposition

If a confidence ring (c-ring, for short) contains a greatest totally ordered subfield, we will call this substructure the "backbone" of the c-ring, and will say it has the backbone property (BBP). The elements of such a backbone can be seen as numerical entities, because they are all totally ordered superfields of **Q**.

Although it is an open question if all c-rings have the BBP, there is an important class of c-rings, the subdirect sums of totally ordered fields, which satisfy BBP. Therefore the discussion of c-rings with BBP covers at least an important part of all models of NC_{12}.

Given the three properties of a c-ring defined below it can be shown that every element of the $[0,1]$-interval can be decomposed into a numerical part and an "interaction" part:

$$\mathbf{c} = b + r \cdot \mathbf{a}$$

where $\mathbf{c}$ is an arbitrary element of the $[0,1]$-interval, b and r are elements of the backbone and $\mathbf{a}$ is a "pure interaction element" of the $[0,1]$-interval, that is, the only bounds by backbone elements are 0 from below and 1 from above.

In order to specify the conditions which entail such a decomposition, we need the following definitions:

Definition: Let C be a c-ring with backbone B. A *cut* of B is a pair (B_1, B_2) of subsets of B with:

1. $B_1 \cap B_2 = \emptyset$
2. $B_1 \cup B_2 = B$
3. $\forall x \in B_1, y \in B_2 : x < y$

Definition: A c-ring C having the BBP is *auto-complete*, if every self-generated cut, i.e., a cut induced by an element of C, has a cut number, i.e., an element of B which generates the given cut.

Definition: A c-ring C is *order-continuous*, if for all $U \subset C$ and $x \in C$ the following holds:

If $U \leq x$, then $\bar{U} \leq x$,

where $\bar{U}$ is the closure of U wrt. the topology induced by all ϵ-balls, ϵ any backbone element greater than 0. In this context, the backbone can be seen as the range of a generalized metric, which then is used to induce a topology in the usual way.

The three properties of a c-ring:

- BBP
- auto-completeness
- order-continuity

then entail the decomposition of a general uncertainty value.

This decomposition result can be interpreted in the following way: a general uncertainty value can be decomposed into a numerical interval ($[b, b + r]$) and an interaction component **a**. If one neglects the interaction information, this implies that uncertainty *in general* can be represented by a numerical interval. This can be used as an argument to propose the use of numerical intervals in the communication of uncertainty. Especially public discussions about forecasts and risks could benefit from a sensible use of intervals, which communicate the estimated reliability of forecasts in an intuitive manner and thus would be very helpful in the decision making process.

Figure 3 shows how a general element **c** of the $[0, 1]$-interval can be thought to be bounded by elements c_* and c^* from the backbone from below and from above.

Uncertainty Decomposition Theorem: If a confidence ring has the BBP, is auto-complete and order-continuous, then all non-backbone elements **c** of the $[0, 1]$-interval can be uniquely represented as:

$$\mathbf{c} = b + r \cdot \mathbf{a}$$

where b and r are elements of the backbone and **a** is an interaction element.

Proof: Let C_* be the lower induced cut and C^* be the upper induced cut generated by **c**, i.e.:

$$C_* = (C_{*,1}, C_{*,2}) = (\{b \in B_{[0,1]} | b < \mathbf{c}\}, B_{[0,1]} \backslash C_{*,1})$$

and

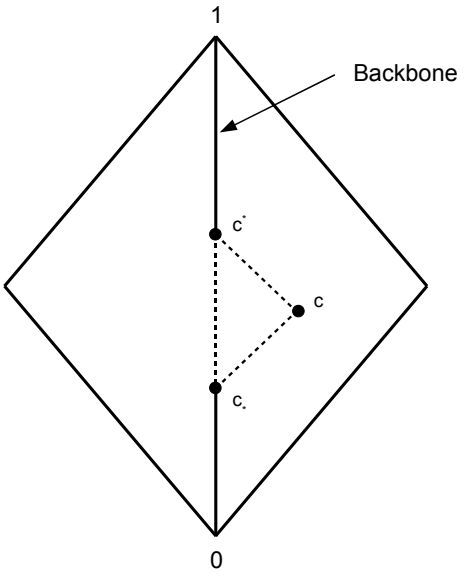

Figure 3: Decomposition of a general uncertainty element.

$$C^* = (C_1^*, C_2^*) = (B_{[0,1]} \backslash C_2^*, \{b \in B_{[0,1]} | b > \mathbf{c}\})$$

The auto-completeness of the confidence ring then ensures that these cuts can be generated by elements of the backbone. Let the corresponding cut numbers be denoted c_* and c^*.

First we will show that $c^* > c_*$. The cut numbers lie in the closures of the cut sets, because otherwise there would be an ϵ-environment of a cut number containing no element of the cut sets. But such a "hole" between the cut sets would violate property 2 of a cut. Order-continuity then results in the following chain of inequalities:

$$c_* \leq \mathbf{c} \leq c^*$$

This immediately yields $c^* \geq c_*$. If $c^* = c_*$, then also $\mathbf{c} = c_*$, i.e. $\mathbf{c}$ would be a backbone element, contrary to the assumption in the theorem.

Now let us define a new element of the confidence ring:

$$\mathbf{a} = (\mathbf{c} - c_*)/(c^* - c_*)$$

The backbone is a field, and thus the difference $c^* - c_*$ is again in the backbone and has an inverse, because the above shown inequality $c^* > c_*$ implies that $c^* - c_* \neq 0$. That is, $\mathbf{a}$ is well-defined.

Next we will show that $\mathbf{a}$ is an interaction element. Let us assume that there is a backbone element $d > 0$ and $\mathbf{a} \geq d$. Then:

$$(\mathbf{c} - c_*)/(c^* - c_*) \geq d$$

$$\mathbf{c} \geq c_* + d \cdot (c^* - c_*)$$

If this inequality would be valid, then c_* would not be the lower cut number of $\mathbf{c}$. Analogously one can show that only 1 is an upper bound from the $[0, 1]$-backbone interval for $\mathbf{a}$. Hence $\mathbf{a}$ is an interaction element.

Now define $b = c_*$ and $r = c^* - c_*$, and the existence of a decomposition follows.

In order to show that the decomposition is unique, we first derive that the lower cut number $b = c_*$ is unique. For this, let us assume that there are two decompositions of c:

$$c = b_1 + r_1 \cdot \mathbf{a_1}$$

and

$$c = b_2 + r_2 \cdot \mathbf{a_2}.$$

By definition, b_1 has the property:

$$b \leq c, b \in B_{[0,1]} \Rightarrow b \leq b_1$$

The same holds for b_2. Thus $b_2 \leq b_1$ and $b_1 \leq b_2$. Hence $b_1 = b_2$.

The same argument works for the upper cut number c^*, which implies that $r = c^* - c_*$ is unique, too. We now have the equation:

$$b + r \cdot \mathbf{a_1} = b + r \cdot \mathbf{a_2}.$$

By subtracting b and dividing by r (remember, that $r > 0$ and all non-zero elements of the backbone are invertible), we finally get:

$$\mathbf{a_1} = \mathbf{a_2},$$

which establishes the uniqueness of the decomposition.

■

6 The Lineage of $\mathbf{NC}_{12}$

The above approach of axiomatizing uncertainty measures extends a line of thinking started by R. T. Cox in 1946 [2]. In [Cox46], based on axioms which should hold for all uncertainty measures, Cox derived a theorem stating that uncertainty measures are essentially probability measures, although his axioms are very different from the axioms of probability theory. A recent exposition of his result can be found in [Jay03].

The application of probability theory to the problem of inductive logic is known as Bayesian inference. Despite its intuitive appeal and many successful applications, it was never considered as a solution to the problem of induction because of technical and philosophical problems. In fact, the 20th century witnessed a strong rejection of probability theory as a theory for induction. Probability theory was developed to describe the randomness of observable events, not the plausibility of unobservable hypotheses. The randomness of events can be seen as an objective property of a physical system, whereas the plausibility of hypotheses is intrinsically subjective, depending on the knowledge of an "observer".

The approach used by Cox was one of the first attempts to justify the use of probabilities as a representation of uncertainty by directly axiomatizing the intuition on uncertainty measures and then *deriving* that uncertainty measures have the same mathematical structure as probability measures. This was a surprising result, given the fact that Cox's axioms look totally different from the Kolmogorov axioms of probability theory. But despite its new and far reaching conclusions, Cox's theorem was not widely acknowledged. This can be attributed to at least two factors: first, it became clear that Cox's derivation of his theorem was not

[2] Glenn Shafer made me aware that a similar approach was already introduced by Sergei Bernstein in 1917, see [SV06] and [Ber17].

complete. The assumptions he made were not sufficient to reach the conclusion in its full generality. This was noted by several authors, and J. Halpern showed in detail where Cox's proof failed by constructing a counterexample in [Hal99]. It was not before 1994 that J.B. Paris completed Cox's proof by introducing a new axiom [Par94]. This axiom closes the holes in Cox's proof, but is very technical in nature. Thus it is not acceptable as an axiom which should hold for all reasonable uncertainty measures. This leads to the second factor contributing to the slow adoption of Cox's result: there is at least one axiom which is too strong to be considered as a general property of uncertainty measures, yet is inherently necessary for the proof approach adopted by Cox. This axiom is the assumption that uncertainty can be measured by one real number. This is a strong structural assumption, implying that the uncertainty values are totally ordered. This prevents, for example, the applicability of Cox's theorem to calculi like Dempster-Shafer theory, which uses two real numbers for the representation of uncertainty.

Additionally, Cox assumes differentiability of functions representing logical connectives. Taken together, these assumptions prevented the applicability of Cox's theorem to calculi like Dempster-Shafer theory, which uses two real numbers for the representation of uncertainty, or fuzzy logic, which uses non-differentiable operations as logical connectives.

The remaining question after the result of J. B. Paris is the following: are there extensions or modifications of the Cox axioms, which are justifiable as general properties of uncertainty measures and which imply a result essentially similar to Cox's theorem? One important step in this direction was taken by S. Arnborg and G. Sjödin. They replaced the axiom introduced by J.B. Paris by a more intuitive statement which they called "Refinability axiom". Furthermore, they dropped the requirement that uncertainty values are real numbers. By this step, they transformed the Cox approach to a genuine algebraic approach, constructing the structure of the domain of uncertainty values and not assuming it. But in order to get the result they wanted, they introduced a total of 16 axioms (when one counts every discernible requirement they formulate as a separate axiom, as we do for our core system), with different degrees of foundational justifiability. Additionally, at a crucial step in their proof they introduce a total order assumption for the domain of uncertainty values, thus restricting the range of their result in a fundamental way.

This was the situation when we entered the development, seeing that Arnborg and Sjödin made a crucial step in the amelioration of the original Cox's approach, but still leaving some major issues open, which have blocked the general applicability of their result. Accordingly, our goal was the following: to devise an axiom system as

minimal as possible, with as weak and as general properties as possible, especially to drop the total order assumption, but still be able to derive a Cox-style result.

In the next section we discuss in detail the evolution of axiom systems which have led to the introduction of NC_{12}.

6.1 The Axiom System of Cox

1. $\mathcal{C} \subseteq \mathbb{R}$ (i.e., confidence values are real numbers)
2. There is a function F: $\Gamma(AB|C) = F(\Gamma(A|BC), \Gamma(B|C))$.
3. There is a function S: $\Gamma(\neg A|B) = S(\Gamma(A|B))$.
4. F, S are twice differentiable.

Cox shows that conditional confidence measures satisfying these axioms are rescalable to probability measures, but his axioms are insufficient to show the full result. This can be seen by analyzing an important method in the proof of Cox, a method one can call "transfer principle" and which is used to transport structure from the Boolean algebra of propositions to the domain of confidence values (in Cox's case the real numbers). But this principle can only be applied if for a given relation on the confidence domain there are a confidence measure and preimages of confidence values which satisfy certain constraints. An example is the Associativity equation for F, which can be reduced to the associativity of the conjunction operator of the Boolean algebra, but only if one finds enough *independent* propositions (see the proof of the associativity of F). Here Cox just assumes that such constrained triples always exist, but that is not ensured by his axioms and indeed one can construct counterexamples, where associativity for F fails. This was first done by J. Halpern in [Hal99]. In order to facilitate the understanding of the innovations which address the problem of missing preimages in the axiom system of Paris, of Arnborg-Sjödin, and NC_{12}, we will now have a closer look at the counterexample constructed by Halpern.

6.1.1 The counterexample of Halpern

Halpern states a theorem which says that there are structures satisfying the axioms of Cox, but F fails to be associative. He proves this by constructing a counterexample, thus showing that Cox's result can not be fixed without a modification of the axiom system.

Theorem: There is a function Bel_0, a finite domain W, and functions S, F, and G satisfying Cox's axioms such that

1. $Bel_0(V|U) \in [0,1]$ for $U \neq \emptyset$
2. $S(x) = 1 - x$
3. $G(x,y) = x + y$
4. F is infinitely differentiable, nondecreasing in each argument in $[0,1]^2$, and strictly increasing in each argument in $(0,1]^2$. Moreover, F is commutative, $F(x,0) = F(0,x) = 0$, and $F(x,1) = F(1,x) = x$.

However, there is no one-to-one function $g : [0,1] \to [0,1]$ satisfying $g(Bel_0(V|U)) \cdot g(Bel_0(U)) = g(Bel_0(VU))$, if $U \neq \emptyset$.

The proof uses a domain with 12 elements, and defines a confidence measure by setting specific values to the elementary events of the finite domain W in a way that the conditions of the theorem are satisfied. However, F is not associative, and the reason is that the finite domain W does not provide enough independent elements in order to apply the transfer principle to all triples of confidence elements, i.e., the triple of confidence values which violates the associativity of F has no preimage of elements in W satisfying the independence constraint. In this specific sense one can say that the world W is *just too small*.

In fact, the following axiom systems try to fix this problem by ensuring that there are "large enough" domains (Paris, Arnborg-Sjödin) or by introducing a "multi-world" framework (NC_{12}), which ensures that there are always enough preimages for constrained triples of confidence values. Principles ensuring such a richness of domains of interest are known as "plenitude principles", occuring also in different areas of philosophy, especially in the context of ontological considerations.

The plenitude principle used by J. Paris is his axiom 5, which directly states that there always will be preimages of confidence triples satisfying certain constraints. This leads to a necessarily infinite, even uncountable, proposition algebra, which is not appealing from a computer science point of view.

Arnborg and Sjödin address this problem by their "Refinability axiom", which in a sense states that proposition algebras can be extended in a certain way, making the world "larger".

In NC_{12} the plenitude principle is embodied in the Extensibility axiom, which states that one can always combine smaller worlds into larger ones. The independence lemma as a direct consequence of the Extensibility axiom then ensures the existence of enough preimages for constrained triples.

So one can see that the core of closing the hole in Cox's proof is to provide a rich enough universe (or "multiverse") of propositions which enables the application of the transfer principle in all relevant cases. The only difference of the following axiom systems in this regard is the degree of "naturalness" or "justifiability" of the introduced plenitude principles.

6.2 The Axiom System of Paris

In [Par94] an axiom system is introduced which is based on Cox's approach, but makes the implicit assumptions of Cox's proof explicit.

1. $\mathcal{C} = [0,1]_{\mathbb{R}}$ ([0,1]-interval of the real numbers)
2. If $A \leq B$ (i.e., $AB = A$), then $\Gamma(B|A) = 1$ and $\Gamma(\neg B|A) = 0$
3. $\Gamma(AB|C) = F(\Gamma(A|BC), \Gamma(B|C))$ for some continuous function F which is strictly increasing (in both arguments) on $(0,1]^2$.
4. $\Gamma(\neg A|B) = S(\Gamma(A|B))$ for some decreasing function S.
5. For any $0 \leq \alpha, \beta, \gamma \leq 1$ and $\epsilon > 0$ there are A_1, A_2, A_3, A_4 with $A_1A_2A_3$ consistent such that each of

$$|\Gamma(A_4|A_1A_2A_3) - \alpha|,\ |\Gamma(A_3|A_1A_2) - \beta|,\ |\Gamma(A_2|A_1) - \gamma|$$

   is less than ϵ.

This last axiom, along with the additional properties of F and S, fills the hole in the proof of Cox, but is hardly intuitive.

Paris-Cox Theorem

Given these axioms, there is a continuous, strictly increasing, surjective function $g : [0,1] \to [0,1]$ such that $\hat{\Gamma} = g \circ \Gamma$ satisfies

$$\begin{aligned} \hat{\Gamma}(\top|\cdot) &= 1 \\ \hat{\Gamma}(A \vee B|C) &= \hat{\Gamma}(A|C) + \hat{\Gamma}(B|C) \qquad \text{if } AB = \bot \\ \hat{\Gamma}(AB|C) &= \hat{\Gamma}(A|BC) \cdot \hat{\Gamma}(B|C) \end{aligned}$$

6.3 The Axiom System of Arnborg and Sjödin

Arnborg and Sjödin replace the fifth axiom of Paris by a more intuitive "Refinability axiom", which consists of three statements:

1. For every confidence space $(\mathbf{U}, \Gamma, \mathcal{C})$, it must be possible to introduce a new subcase B of a non-false proposition A with confidence value v given to $\Gamma(B|A)$.

2. If two new subcases B and B' of a proposition A are defined, they can be specified to be independent, i.e., $\Gamma(B|B'A) = \Gamma(B|A)$ and $\Gamma(B'|BA) = \Gamma(B'|A)$.

3. For two confidence values v, w such that $v < S(w)$, it should be possible to define two new subcases C, C' of any non-false proposition A such that $v = \Gamma(C|A)$, $w = \Gamma(C'|A)$ and $\Gamma(CC'|A) = \perp\!\!\!\perp$.

Arnborg and Sjödin introduce functions F and S like Cox, but additionally a *partial* function G (only defined for confidence values v, w with $v \leq S(w)$) such that:

$$\Gamma(A \vee B|C) = G(\Gamma(A|C), \Gamma(B - A|C)).$$

Furthermore, in a later part of their paper, they introduce a total order assumption for the domain of confidence values. Their other axioms are very similar to the axioms of Paris, but they take the important step to drop the real value assumption and derive the algebraic structure of the domain of confidence values from their axioms.

Their main result is the following: given their axioms, the domain of confidence values can be embedded in a totally ordered field, where multiplication and addition are extensions of F and G. Analyzing their proof, we find that the construction of a field from a ring will fail if one does not assume a total order on $\mathcal{C}$. In lemma 13 of [AS01] they state that the ring they have constructed is a totally ordered integral domain, i.e. a ring without zero divisors (an example of zero divisors in a function ring is depicted in figure 4). Then they use a theorem from S. MacLane and G. Birkhoff in [MB67] which states that every totally ordered integral domain can be embedded in a totally ordered field. But this will not work in the case of partial order because without the total order assumption one cannot prove that

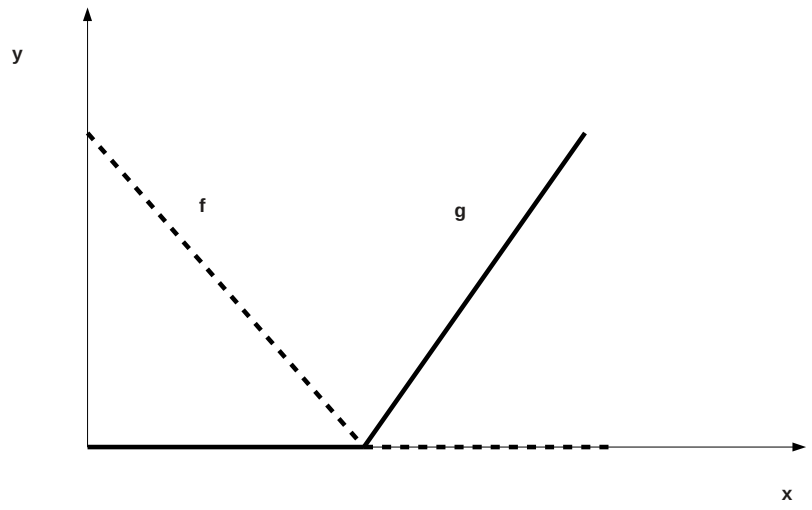

Figure 4: Zero divisors in a function ring: $f \not\equiv 0, g \not\equiv 0, f \cdot g \equiv 0$.

the constructed ring will not contain zero divisors. So, lemma 13 of [AS01] cannot be transferred to the partial order case, which blocks the application of the MacLane-Birkhoff theorem. This is an interesting example of the interplay between order properties and algebraic properties: a total order assumption has strong algebraic implications, while partial order has not. Accordingly, order properties and algebraic properties cannot, as one might have hoped, be treated separately.

7 Relations to existing Uncertainty Calculi

Today, there exist many approaches for dealing with uncertainty, for example lower probabilities, which have only partially ordered uncertainty values or non-monotonic logic, which can be interpreted as using infinitesimal probabilities. In the following, we try to analyze these calculi in the light of our results.

7.1 Lower Probabilities

The problem of dealing with "imprecise" probabilities has led to the development of calculi known under the common name "lower probabilities". The main distinction

from the probability calculus is that the uncertainty of a proposition is judged by *two* numbers instead of one. Accordingly, there are two functions mapping the elements of a proposition algebra to $[0, 1]$, the *lower probability* P_* and the *upper probability* P^*. The most general notion of a lower probability is defined wrt. a set of probability distributions $\mathcal{P}$ (see, for example, [Hal03]):

$$P^*(A) = \sup_{P \in \mathcal{P}} P(A) \quad \text{and} \quad P_*(A) = \inf_{P \in \mathcal{P}} P(A)\,.$$

One can show that lower and upper probabilities satisfy the following inequalities if A and B are disjoint:

$$P_*(A \cup B) \geq P_*(A) + P_*(B) \quad \text{and} \quad P^*(A \cup B) \leq P^*(A) + P^*(B)\,.$$

These properties are called super-additivity and sub-additivity, respectively. Furthermore, lower and upper probability are connected via the following relations:

$$P_*(A) \leq P^*(A) \quad \text{and} \quad P^*(A) = 1 - P_*(\bar{A})\,.$$

The inequality says that lower and upper probabilities can be seen as defining an interval, thus making lower and upper probabilities an uncertainty calculus having a partially ordered domain of uncertainty values. The equation implies that from both uncertainty values, upper and lower probability, of a proposition one can derive the upper and lower probabilities of its negation. Hence lower and upper probabilities together satisfy axiom Not.

An application of our results to the analysis of lower probabilities is now the following: even if the domain of uncertainty values is only partially ordered, which is possible according to NC_{12}, there exists a function G which relates the uncertainty value of a disjunction of disjoint propositions and the uncertainty values of the single propositions by an equation, and not only by an inequality. If no such function G exists for an uncertainty calculus, it must violate at least one of the axioms Not, And_1, or And_2 (we assume that the infrastructure axioms are satisfied). Now, because lower probabilities satisfy axiom Not, they must violate And_1 or And_2. This implies that there cannot be any definition of conditioning for lower probabilities which satisfies And_1 and And_2. Seeing And_1 and And_2 as essential conditions for not losing relevant information, this may explain why the definition of conditioning for lower probabilities has turned out to be such a hard problem, which is still the topic of ongoing research.

This conclusion is also valid for Dempster-Shafer theory, which can be seen as lower and upper probabilities satisfying additional constraints. Accordingly, there are several proposals for conditioning in DS-theory, each having its own advantages and disadvantages. By the above analysis, this is not a transitory state until the "right" conditioning rule has been found, but a fundamental obstacle which cannot be resolved within the frame of DS-theory.

7.2 Dempster-Shafer Theory

In 1967 A. P. Dempster introduced a generalization of probability theory, which was later extended by G. Shafer [Dem67, Sha76]. An essential difference to probability theory is that uncertainty is now represented by two functions, which are called "*belief*" and "*plausibility*" (it can be argued that "*possibility*" instead of "plausibility" is a better name for the second function, but the use of "plausibility" is established today).

One way to define belief functions is by introducing a "mass function" or "basic belief assignment". Let $\mathbf{U}$ be a finite Boolean algebra (which is called a "frame of discernment" by Shafer), then a mass function is defined as a mapping m from $\mathbf{U}$ to the real [0,1]-interval satisfying the following conditions:

1. $m(\bot) = 0$
2. $\sum_{A \in \mathbf{U}} = 1$

So the "belief masses" have to sum up to 1, as in probability theory, but now the masses can be distributed over *all* elements (except the bottom element) of the Boolean algebra, whereas in probability theory the masses have to be assigned only to the *atoms* of the Boolean algebra.

The intuition behind this is the following: a mass assigned to a non-atomic element of the Boolean algebra represents ignorance with regard to the distribution of this mass to the atoms lying below the non-atomic element. Consequently, the mass function which assigns 1 to the top element of the Boolean algebra and 0 to all other elements represents maximal ignorance and is called the *vacuous* mass function.

The mass function can now be used to define a belief function Bel:

$$Bel(A) = \sum_{B \leq A} m(B)$$

So to obtain the total belief committed to A, one adds to the mass of A all the masses assigned to propositions which lie below A.

Belief functions are super-additive, in contrast to probability functions, which are additive:

$$Bel(A \vee B) \geq Bel(A) + Bel(B), \text{ if } A \wedge B = \bot$$

We will now investigate the relationship of DS-theory to confidence theory. A basic question in this context is the following:

Can all belief functions be represented by confidence measures?

A confidence measure represents a belief function if it is decomposable in the sense introduced in chapter 5, the backbone field is $\mathbb{R}$ and the function which assigns to all propositions the lower bound of the numerical interval resulting from the decomposition coincides with the given belief function.

Using this definition of representation, the answer to the above question is yes, at least in the case of finite proposition algebras. In order to prove this, the following notion will play a central role:

Definition: An *interference family of order* n is a set of n (multivariate) real-valued functions $f_1, .., f_n$ with the following properties:

1. The range of $\sum_{i \in I} f_i$ is [0,1], where I is any nonempty, proper subset of the index set $\{1, .., n\}$.
2. $\sum_{i=1}^{n} f_i = 1$

That is, only if we add all n functions the range collapses to the point set $\{1\}$, all other sums where at least one function is missing have still the full range [0,1]. Such families can now be used to construct a confidence function representing a given belief function Bel. Below we will construct interference families for all orders.

As a c-ring we choose the function ring over the real [0,1]-interval with countably many variables, the constant 1-function is the top confidence value and the constant 0-function is the bottom confidence value. The order is defined as follows:

$$f \leq g \;\Leftrightarrow\; f(\mathbf{x}) < g(\mathbf{x}) \text{ for all } \mathbf{x} \text{ or } f \equiv g$$

We will now construct a confidence function over this c-ring which represents the given belief function Bel.

Let the *order of a proposition* be the number of atoms it subsumes. In a first step we assign to an arbitrary element A of the (finite) Boolean algebra $\mathbf{U}$ of order j an interference family F_A of order j. Next we note that confidence measures are additive, thus it suffices to define the confidence value for all atoms. So let A_i be an atom and $\mathcal{F}(A_i)$ the filter generated by the atom A_i, i.e., the set of all propositions which subsume A_i as an atom. For every element A of $\mathcal{F}(A_i)$ we pick an element of the interference family we have assigned to A. Let us denote this interference function by $f_A^{(A_i)}$.

We now define a confidence measure Γ_{Bel} on the Boolean algebra $\mathbf{U}$ by defining its values on the atoms of $\mathbf{U}$:

$$\Gamma_{Bel}(A_i) = \sum_{A \in \mathcal{F}(A_i)} m_{Bel}(A) \cdot f_A^{(A_i)}$$

where m_{Bel} is the mass function corresponding to the given belief function Bel (see, e.g., [Sha76], theorem 2.2, p. 39).

This construction ensures that the lower bounds of the intervals defined by Γ_{Bel} are exactly the sums of the masses for which the interference families have collapsed, and this happens for those propositions which are subsumed by a given proposition, i.e.:

$$\Gamma_{Bel,*}(A) = \sum_{B \leq A} m_{Bel}(B)$$

From the definition of a belief function it follows that $\Gamma_{Bel,*}$ equals the given belief function Bel, hence Γ_{Bel} is a representation of Bel.

7.2.1 Interference Families

Finally, we have to show that there are interference families of all orders. This is established by the following theorem:

Theorem: Let $\{\alpha_1, .., \alpha_k\}$ be a set of k variables ranging over [0,1]. Then the functions $(1 \leq j \leq 2^k)$

$$^{(k)}f^{(j)} = f_1 \cdot f_2 \cdot .. \cdot f_k,$$

where $f_i = \alpha_i$ or $f_i = 1 - \alpha_i$, constitute an interference family of order $n = 2^k$.

Proof: By induction.

k=1: It is straightforward to check that the function set $\{\alpha_1, 1-\alpha_1\}$ is an interference family of order 2.

k+1 step: The functions ${}^{(k+1)}f^{(j)}$ can be written as:

$$ {}^{(k+1)}f^{(j)} = \alpha_{k+1} \cdot \sum_{j_1 \in I_1} {}^{(k)}f^{(j_1)} + (1-\alpha_{k+1}) \cdot \sum_{j_2 \in I_2} {}^{(k)}f^{(j_2)} $$

If both, I_1 and I_2, are complete index sets, i.e. $I_1 = I_2 = \{1, .., 2^k\}$, then, according to induction assumption, both sums collapse to 1, leaving $\alpha_{k+1} + (1-\alpha_{k+1}) = 1$. Thus the collapse property is valid for $k+1$, too.

If at least one index set is incomplete, the corresponding sum function has the full range [0,1], by induction assumption. If I_1 is incomplete, we set α_{k+1} to 1 in order to see that ${}^{(k+1)}f^{(j)}$ has full range, too. If I_2 is incomplete, then we choose $\alpha_{k+1} = 0$. If both, I_1 and I_2, are incomplete, then both choices, $\alpha_{k+1} = 0$ and $\alpha_{k+1} = 1$ show that ${}^{(k+1)}f^{(j)}$ has still full range. Hence the first property of an interference family is valid for $k+1$.

■

If n is not a power of two, then we take the interference family with the least k so that $n < 2^k$. Set $g_j = {}^{(k)}f^{(j)}$ for all $j < n$ and $g_n = \sum_{j=n}^{2^k} {}^{(k)}f^{(j)}$. Thus there is an interference family for all orders.

7.3 Non-monotonic Logic

A non-monotonic logic extends classical logic with a framework of "belief revision", i.e. conclusions derived at one point can be retracted at a later point. Non-monotonic logic can be seen as defining a hierarchy of "default assumptions", which are assumed valid until observed evidence directly contradicts them. If this happens, a revision process is executed, which incorporates the new evidence and eliminates contradictions while trying to preserve as much as possible from the old knowledge state. Now, as for example Lehman and Magidor have observed in [LM92], one can formalize default expressions of the type "if A then typically B" as "the probability of B given A is very high", where "very high" is equated to $1-\epsilon$, for *infinitesimal* ϵ. This can be modeled by a generalized probability algebra using the $[0,1]$-interval of *hyperreal* numbers as a domain of uncertainty values.

8 Conclusions

Despite many attempts, there is still no consensus on basic questions concerning uncertainty and the foundations of inductive logic. In [AS01], Arnborg and Sjödin note that reaching a consensus is not only a foundational issue but is also important outside the ivory tower: designers of complex systems struggle with difficult compatibility problems when they plan to integrate system components which happen to use different ways to describe uncertainty [Zad97, Wal96].

This thesis tries to contribute to the debate on uncertainty by discerning ontologically different types of uncertainty and introducing an axiomatic core system for uncertainty measures with the explicit aim not to prejudice structural properties of the domain of uncertainty values, but to derive them from basic assumptions.

The main result characterizes uncertainty values as elements of the [0,1]-interval of a partially ordered ring, including the Kolmogorov axioms of probability theory as special case, but allowing uncertainty domains having uncomparable or infinitesimal elements as well.

9 Effective Learning Systems

In the previous part we have analyzed the problem of representing and processing uncertainty with regard to a *given set of models*. This was the first question posed in the introduction. Now we shortly want to discuss the second question of the introduction:

What set of possible models of the environment should the learning system consider?

To answer this question, it is necessary to explore the notion of "all possible models" from a mathematical and computational point of view, and discuss the question of effective learnability in the context of such generic model spaces. The last sentence intentionally uses the plural form of model space, because we will see that the notion of "all possible models" cannot be defined in an absolute sense, but only with regard to a reference proof system. This dependence can be used to establish a relationship between the *time* complexity of the percept-generating environment and the *logical* complexity of the algorithmic learning system, thus shedding new light on the undecidability of a general approach to algorithmic learning developed by R. J. Solomonoff in the 1960s [Sol64a, Sol64b, LV08].

In order to focus on the analysis of the model space, we now will use the framework of Bayesian inference to handle uncertainty. Although this does not make use of the full generality of the uncertainty calculus developed in the first part of this thesis, Bayesian inference is a model of the axiom system NC_{12} - as we have already noted in the discussion of the ring theorem - and therefore justified by this axiom system as an adequate calculus to represent and process uncertainty.

9.1 Algorithmic Ontology: Programs as Generators

Ontology is a part of philosophy which tries to define what exists, or, more specifically, what *possibly* could exist. In the realm of mathematics, this question leads to the set existence problem, which is (partially) answered by various set theories, most commonly by using the axiom system **ZFC**, Zermelo-Fraenkel set theory with the axiom of choice. But in the realm of computer science, existence has to be *effective* existence, i.e. the domain of interest and its operations must have effective representations.

For this reason the objects we consider are programs executed by a fixed universal Turing machine U having a one-way read-only input tape, some work tapes, and

a one-way write-only output tape (such Turing machines are called monotone), which will be the reference machine for all that follows. The choice of the specific universal Turing machine has as effect only a constant factor on the space complexity and at most a logarithmic factor on the time complexity of a program [AB09]. These effects are small enough to be neglected in the following foundational considerations, but in the context of alternative models of computation, like cellular automata on infinite grids, the choice of the reference machine may become an important issue. The program strings are chosen to be prefix-free, i.e. no program string is the prefix of another program string. This is advantageous from a coding point of view (enabling, for example, the application of Kraft's inequality $\sum_p 2^{-length(p)} \leq 1$), and does not restrict universality [LV08].

A program p (represented as finite binary string) is a generator of a possible world, if it outputs an infinite stream of bits when executed by U. Unfortunately, it is not decidable whether a given program p has this well-definedness property. This is the reason why the general approach to induction introduced by R. J. Solomonoff is incomputable: the inference process uses the whole set of programs (program space) as possible generators, even the programs which are not well-defined in the above sense.

This results in the following dilemma: either one restricts the model space to a decidable set of well-defined programs, which leads to an effective inference process but ignores possibly meaningful programs, or one keeps all well-defined programs, but at the price of necessarily keeping ill-defined programs as well, risking the incomputability of the inference process. However, in the following we propose an approach which tries to mitigate this dilemma by reducing the question of learnability to the question of provability.

9.2 Learning Systems

The following discussion is based on [ZC12]. Here we give only an introduction into the basic notions and results, and put them into perspective.

First the notion of a *probabilistic learning system* is introduced, which takes a finite string of observed bits (the percept string) as input and produces a probabilistic prediction for the next bit as output:

Definition: A *probabilistic learning system* is a function

$$\Lambda : \{0,1\}^* \times \{0,1\} \to [0,1]_{\mathbf{Q}}, \quad \text{with } \Lambda(x,0) + \Lambda(x,1) = 1 \text{ for all } x \in \{0,1\}^*.$$

Λ is an *effective* probabilistic learning system if Λ is a total recursive function. Rational numbers are used as probability values, because real numbers cannot be used directly in a context of computability questions, but have to be dealt with by effective approximations [Wei00]. This would increase the complexity of the definitions significantly, and is not necessary for a first understanding of the fundamental relationship between learnability and provability. Also only deterministic generators leading to one definite observable bit sequence are treated, but a generalization to real valued probabilities and probabilistic generators should be possible.

Next the prediction horizon of Λ is extended by feeding it with its own predictions. This leads to a learning system $\Lambda^{(k)}$ which makes probabilistic predictions for the next k bits (xy is the concatenation of string x and y):

$$\Lambda^{(1)} = \Lambda,$$

$$\Lambda^{(k+1)}(x, y1) = \Lambda^{(k)}(x, y) \cdot \Lambda(xy, 1), \quad x \in \{0,1\}^*, y \in \{0,1\}^k,$$

$$\Lambda^{(k+1)}(x, y0) = \Lambda^{(k)}(x, y) \cdot \Lambda(xy, 0).$$

Finally, the learnability of an infinite bit sequence s is defined ($s_{i:j}$ is the subsequence of s starting with bit i and ending with bit j):

Definition: An infinite bit sequence s is learnable in the limit by the probabilistic learning system Λ, if for all $\epsilon > 0$ there is an n_0 so that for all $n \geq n_0$ and all $k \geq 1$:

$$\Lambda^{(k)}(s_{1:n}, s_{n+1:n+k}) > 1 - \epsilon.$$

Based on this notion of learnability, a proof-driven learning system is introduced in [ZC12]. The basic idea is to take a background theory Σ from the foundations of mathematics, like the ones investigated in reverse mathematics [Sim09], and to use the provably total recursive functions wrt. Σ to construct a fast growing guard function to schedule the learning process and to ensure that the learning system is effective. The resulting learning system is denoted by $\Lambda(\Sigma)$.

The generator-predictor theorem, a theorem characterizing the learnable bit sequences of proof-driven learning systems, states that an infinite sequence of bits is learnable if the axiom system proves the totality of a recursive function which dominates the time function of the bit sequence generating process:

Generator-Predictor Theorem: Let Σ be an admissible logic frame and s an infinite bit sequence. s is learnable by the effective probabilistic learning system

$\Lambda(\Sigma)$, if Σ proves the totality of a recursive dominator of a generator function G_p for at least one program $p \in [s]$,

where G_p is the *generator function* of the program p, i.e., the function counting the number of executed steps in order to generate the first n bits, and $[s]$ denotes the set of all programs generating the bit sequence s.

This result establishes a tight connection between learnability and provability, thus reducing the question of what can be effectively learned to the foundational questions of mathematics with regard to set existence axioms. Results of reverse mathematics are used to illustrate the implications of the generator-predictor theorem by connecting a hierarchy of axiom systems with increasing logical strength to fast growing functions.

For example, if $\Sigma_{PA} = (FOL, PA)$ (First order logic Peano Arithmetic), then every program with a generator function which is dominated by a provably total recursive function wrt. PA can be learned by $\Lambda(\Sigma_{PA})$. As the Ackermann-function is provably total in PA, this is already a pretty large set, and PA allows totality proofs of functions which grow much faster than the Ackermann-function.

9.3 Synchronous Learning Framework

A closer look on real world incremental learning situations, where both, the environment and the learning system, are not suspended while the other one is performing its transitions resp. computations, leads to the following notion of *synchrony* of a bit sequence s:

Definition: s is *synchronous* if $\limsup_{n\to\infty} \frac{G_p(n)}{n} < \infty$ for at least one $p \in [s]$.

Synchrony entails that the time scales of the learning system and the environment are coupled, that they cannot ultimately drift apart. As long as one assumes not a malicious environment, synchrony seems to be a natural property. Such a setting for learning could be called a synchronous learning framework, in contrast to the above considered learning frameworks, which could be classified as asynchronous.

In order to learn in the case of synchrony, it suffices to prove that n^2 is total, because n^2 is a dominator of every generator function satisfying the synchrony condition. Thus, a much weaker background theory than, e.g. Peano Arithmetic would suffice for an effective learning system to learn all synchronously generated bit sequences. In fact, because RCA_0 – the weakest of the five standard axiom systems considered in reverse mathematics – proves the totality of all primitive recursive functions, $\Lambda(RCA_0)$ is a perfect learning system in a synchronous world.

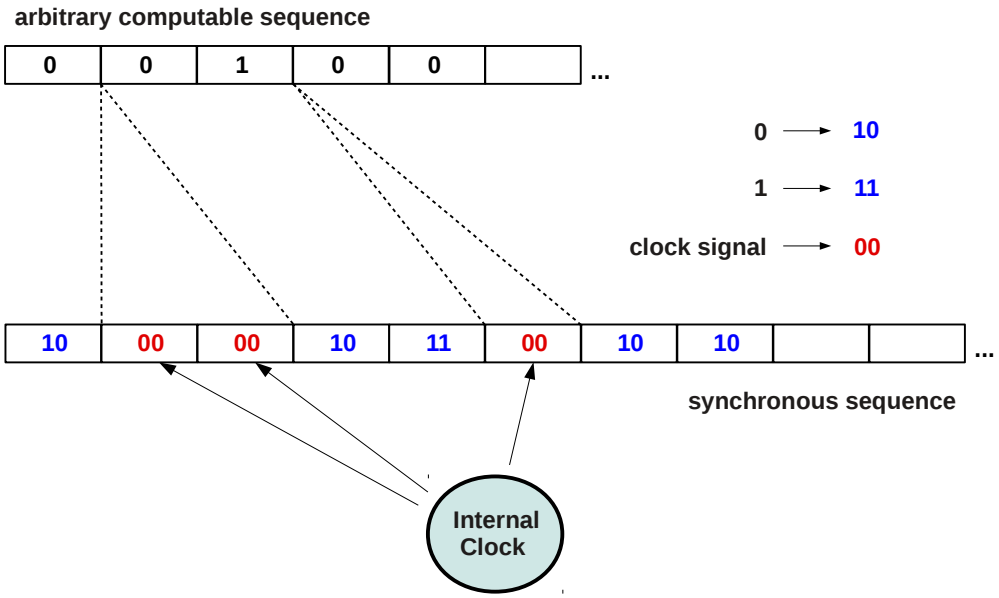

Figure 5: Clockification: using an internal clock transforms all computable bit sequences into synchronous bit sequences.

9.4 Conclusions

The generator-predictor theorem establishes a natural perspective on the effective core of Solomonoff induction by shedding new light on the cause of the incomputability of the non-relativized Solomonoff induction, instead of directly introducing specific resource constraints in order to achieve computability, like this is done, for example, in [Hut05] for the AIξ learning system. This shifts the questions related to learnability to questions related to provability, and therefore into the realm of the foundations of mathematics.

The problem of universal induction in the synchronous learning framework, however, is intrinsically effective, and the focus of future research in a synchronous framework can be on *efficiency* questions. In fact, the source of incomputability in the asynchronous learning framework can be traced back to the fact that the learning system does not know how much time the generator process has "in-

vested" in order to produce the next bit. An extension of a bit sequence s by inserting "clock signals" (coding a clock signal by "00" and output bits by "10" and "11") marking the passing of time would transform every sequence s into a synchronous one, thus eliminating the incomputability of Solomonoff induction. So the synchronous learning framework seems to be perfectly suited for studying the problem of universal induction from a computational point of view.

References

[AB09] S. Arora and B. Barak. *Complexity Theory: A Modern Approach*. Cambridge University Press, 2009.

[AS01] S. Arnborg and G. Sjödin. What is the plausibility of probability? *Preprint, Nada, KTH*, 2001.

[Ber17] Sergei N. Bernstein. On the axiomatic foundation of the theory of probability (in Russian). *Communications of the Kharkiv mathematical society*, 15:209–274, 1917.

[Ber85] James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, second edition, 1985.

[BS94] Jose M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. Wiley, 1994.

[Cha03] I. Chajda. Lattices and semilattices having an antitone involution in every upper interval. *Comment. Math. Univ. Carolinae*, 44(4):577–585, 2003.

[Con76] John H. Conway. *On Numbers and Games*. Academic Press, 1976.

[Cox46] R. T. Cox. Probability, frequency, and reasonable expectation. *Am. Jour. Phys.*, 14:1–13, 1946.

[Dem67] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2):325–339, 1967.

[DP88] D. Dubois and H. Prade. *Possibility Theory*. Plenum Press, New York, 1988.

[Dub06] D. Dubois. Possibility theory and statistical reasoning. *Computational Statistics & Data Analysis*, 51(1):47–69, 2006.

[Gär92] P. Gärdenfors, editor. *Belief Revision*. Cambridge University Press, 1992.

[Gin87] M. Ginsberg, editor. *Readings in Nonmonotonic Reasoning*. Morgan Kauffman, Los Altos, CA, 1987.

[Gon86] Harry Gonshor. *An Introduction to Surreal Numbers*. Cambridge University Press, 1986.

[Hal99] J. Halpern. A counterexample to theorems of Cox and Fine. *Journal of A.I. Research*, 10:76–85, 1999.

[Hal03] J. Halpern. *Reasoning about Uncertainty*. MIT Press, 2003.

[HSP09] F. Huber and C. Schmidt-Petri, editors. *Degrees of Belief*. Springer, 2009.

[Hut05] M. Hutter. *Universal Artificial Intelligence*. Springer, 2005.

[HW96] U. Hebisch and H. J. Weinert. *Semirings and Semifields*, pages 425–462. Handbook of Algebra, Vol. 1. Elsevier, 1996.

[Jay03] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[Kni21] F. Knight. *Risk, Uncertainty, and Profit*. Houghton Mifflin Company, 1921.

[Knu74] Donald E. Knuth. *Surreal Numbers*. Addison-Wesley, 1974.

[LM92] D. Lehman and M. Magidor. What does a conditional knowledge base entail? *Artificial Intelligence*, 55(1):1–60, 1992.

[LV08] Ming Li and Paul M. B. Vitányi. *An introduction to Kolmogorov complexity and its applications (3. ed.)*. Graduate texts in computer science. Springer, 2008.

[MB67] S. MacLane and G. Birkhoff. *Algebra*. The MacMillan Company, 1967.

[Par94] J. B. Paris. *The Uncertain Reasoner's Companion*. Cambridge University Press, 1994.

[Pea00] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[PR08] R. Padmanabhan and S. Rudeanu. *Axioms for lattices and boolean algebras*. World Scientific, 2008.

[Sch09] J. Schmidhuber. Ultimate cognition à la Gödel. *Cognitive Computation*, 1(2):177–193, 2009.

[SF02] K. Sentz and S. Ferson. Combination of evidence in Dempster-Shafer theory. *Sandia National Laboratories Report 2002-0835*, 2002.

[Sha76] G. Shafer. *Mathematical Theory of Evidence*. Princeton University Press, 1976.

[Sim09] S. G. Simpson. *Subsystems of Second Order Arithmetic (2. ed.)*. Cambridge University Press, 2009.

[Sol64a] R. Solomonoff. A formal theory of inductive inference, part I. *Information and Control*, 7(1):1–22, 1964.

[Sol64b] R. Solomonoff. A formal theory of inductive inference, part II. *Information and Control*, 7(2):224–254, 1964.

[Spo99] W. Spohn. Ranking functions, AGM style. In B. Hansson, S. Halldén, N.-E. Sahlin, and W. Rabinowicz, editors, *Internet Festschrift for Peter Gärdenfors, http://www.lucs.lu.se/spinning*, Lund, 1999.

[Spo09] W. Spohn. A survey of ranking theory. In F. Huber and C. Schmidt-Petri, editors, *Degrees of Belief*. Springer, 2009.

[SV06] G. Shafer and V. Vovk. The sources of Kolmogorov's Grundbegriffe. *Statistical Science*, 21(1):70–98, 2006.

[Wal96] P. Walley. Measures of uncertainty in expert systems. *Artificial Intelligence*, 83:1–58, 1996.

[Wei00] K. Weihrauch. *Computable analysis*. Springer, 2000.

[Zad97] L. A. Zadeh. The roles of fuzzy logic and soft computing in the conception, design and deployment of intelligent systems. *Lecture Notes in Computer Science*, 1198:183–210, 1997.

[ZC11] J. Zimmermann and A. B. Cremers. The Quest for Uncertainty. In Cristian S. Calude, Grzegorz Rozenberg, and Arto Salomaa, editors, *Rainbow of Computer Science*, volume 6570 of *Lecture Notes in Computer Science*. Springer, 2011.

[ZC12] J. Zimmermann and A. B. Cremers. Making Solomonoff induction effective or you can learn what you can bound. In S. B. Cooper, A. Dawar, and B. Löwe, editors, *How the World Computes*, volume 7318 of *Lecture Notes in Computer Science*. Springer, 2012.