A Logic for Inductive Probabilistic Reasoning

Manfred Jaeger
Department for Computer Science, Aalborg University
Fredrik Bajers Vej 7E, DK-9220 Aalborg Ø
jaeger@cs.aau.dk

Abstract

Inductive probabilistic reasoning is understood as the application of inference patterns that use statistical background information to assign (subjective) probabilities to single events. The simplest such inference pattern is direct inference: from "70% of As are Bs" and "a is an A" infer that a is a B with probability 0.7. Direct inference is generalized by Jeffrey's rule and the principle of cross-entropy minimization. To adequately formalize inductive probabilistic reasoning is an interesting topic for artificial intelligence, as an autonomous system acting in a complex environment may have to base its actions on a probabilistic model of its environment, and the probabilities needed to form this model can often be obtained by combining statistical background information with particular observations made, i.e. by inductive probabilistic reasoning.

In this paper a formal framework for inductive probabilistic reasoning is developed: syntactically it consists of an extension of the language of first-order predicate logic that allows to express statements about both statistical and subjective probabilities. Semantics for this representation language are developed that give rise to two distinct entailment relations: a relation $\models$ that models strict, probabilistically valid, inferences, and a relation $\mid\!\approx$ that models inductive probabilistic inferences. The inductive entailment relation is obtained by implementing cross-entropy minimization in a preferred model semantics. A main objective of our approach is to ensure that for both entailment relations complete proof systems exist. This is achieved by allowing probability distributions in our semantic models that use non-standard probability values. A number of results are presented that show that in several important aspects the resulting logic behaves just like a logic based on real-valued probabilities alone.

1 Introduction

1.1 Inductive Probabilistic Reasoning

Probabilities come in two kinds: as statistical probabilities that describe relative frequencies, and as subjective probabilities that describe degrees of belief. To both kinds of probabilities the same rules of probability calculus apply, and notwithstanding a long and heated philosophical controversy over what constitutes the proper meaning of probability (de Finetti 1937, von Mises 1951, Savage 1954, Jaynes 1978), few conceptual difficulties arise when we deal with them one at a time.

However, in commonsense or inductive reasoning one often wants to use both subjective and statistical probabilities simultaneously in order to infer new probabilities of interest. The simplest example of such a reasoning pattern is that of *direct inference* (Reichenbach 1949, §72),(Carnap 1950, §94), illustrated by the following example: from

> 2.7% of drivers whose annual mileage is between 10,000 and 20,000 miles will be involved in an accident within the next year (1)

and

> Jones is a driver whose annual mileage is between 10,000 and 20,000 miles (2)

infer

> The probability that Jones will be involved in an accident within the next year is 0.027. (3)

The percentage 2.7 in (1) is a statistical probability: the probability that a driver randomly selected from the set of all drivers with an annual mileage between 10,000 and 20,000 will be involved in an accident. The probability in (3), on the other hand, is attached to a proposition that, in fact, is either true or false. It describes a state of knowledge or belief, for which reason we call it a subjective probability. [1]

Clearly, the direct inference pattern is very pervasive: not only does an insurance company make (implicit) use of it in its computation of the rate it is willing to offer a customer, it also underlies some of the most casual commonsense reasoning ("In very few soccer matches did a team that was trailing 0:2 at the end of the first half still win the game. My team is just trailing 0:2 at halftime. Too bad".), as well as the use of probabilistic expert systems. Take a medical diagnosis system implemented by a Bayesian network (Pearl 1988, Jensen 2001), for instance: the distribution encoded in the network (whether specified by an expert or learned from data) is a statistical distribution describing relative frequencies in a large

[1] Other names for this type of probability are "probability of the single case"(Reichenbach 1949), "probability$_1$"(Carnap 1950), "propositional probability"(Bacchus 1990*b*).

number of past cases. When using the system for the diagnosis of patient Jones, the symptoms that Jones exhibits are entered as evidence, and the (statistical) probabilities of various diseases conditioned on this evidence are identified with the probability of Jones having each of these diseases.

Direct inference works when for some reference class C and predicate P we are given the statistical probability of P in C, and for some singular object e all we know is that e belongs to C. If we have more information than that, direct inference may no longer work: assume in addition to (1) and (2) that

> 3.1% of drivers whose annual mileage is between 15,000 and 25,000 miles will be involved in an accident within the next year (4)

and

> Jones is a driver whose annual mileage is between 15,000 and 25,000 miles . (5)

Now direct inference can be applied either to (1) and (2), or to (4) and (5), yielding the two conflicting conclusions that the probability of Jones having an accident is 0.027 and 0.031. Of course, from (1),(2), (4), and (5) we would infer neither, and instead ask for the percentage of drivers with an annual mileage between 15,000 and 20,000 that are involved in an accident. This number, however, may be unavailable, in which case direct inference will not allow us to derive any probability bounds for Jones getting into an accident. This changes if, at least, we know that

> Between 2.7% and 3.1% of drivers whose annual mileage is between 15,000 and 20,000 miles will be involved in an accident within the next year. (6)

From (1),(2), and (4)-(6) we will at least infer that the probability of Jones having an accident lies between 0.027 and 0.031. This no longer is direct inference proper, but a slight generalization thereof.

In this paper we will be concerned with inductive probabilistic reasoning as a very broad generalization of direct inference. By inductive probabilistic reasoning, for the purpose of this paper, we mean the type of inference where statistical background information is used to refine already existing, partially defined subjective probability assessments (we identify a categorical statement like (2) or (5) with the probability assessment: "with probability 1 is Jones a driver whose…"). Thus, we here take a fairly narrow view of inductive probabilistic reasoning, and, for instance, do not consider statistical inferences of the following kind: from the facts that the individuals $jones_1, jones_2, \ldots, jones_{100}$ are drivers, and that $jones_1, \ldots, jones_{30}$ drive less and $jones_{31}, \ldots, jones_{100}$ more than 15,000 miles annually, infer that 30% of drivers drive less than 15,000 miles. Generally speaking, we are aiming at making inferences only in the direction from statistical to subjective probabilities, not from single-case observations to statistical probabilities.

Problems of inductive probabilistic reasoning that go beyond the scope of direct inference are obtained when the subjective input-probabilities do not express certainties:

> With probability 0.6 is Jones a driver whose annual mileage is between 10,000 and 20,000 miles. (7)

What are we going to infer from (7) and the statistical probability (1) about the probability of Jones getting into an accident? There do not seem to be any sound arguments to derive a unique value for this probability; however, $0.6 \cdot 0.027 = 0.0162$ appears to be a sensible lower bound. Now take the subjective input probabilities

> With probability 0.6 is Jones's annual mileage between 10,000 and 20,000 miles, and with probability 0.8 between 15,000 and 25,000 miles. (8)

Clearly, it's getting more and more difficult to find the right formal rules that extend the direct inference principle to such general inputs.

In the guise of inductive probabilistic reasoning as we understand it, these generalized problems seem to have received little attention in the literature. However, the mathematical structure of the task we have set ourselves is essentially the same as that of *probability updating*: in probability updating we are given a *prior* (usually subjective) probability distribution representing a state of knowledge at some time t, together with new information in the form of categorical statements or probability values; desired is a new *posterior* distribution describing our knowledge at time $t + 1$, with the new information taken into account. A formal correspondence between the two problems is established by identifying the statistical and subjective probability distributions in inductive probabilistic inference with the prior and posterior probability distribution, respectively, in probability updating.

The close relation between the two problems extends beyond the formal similarity, however: interpreting the statistical probability distribution as a canonical prior (subjective) distribution, we can view inductive probabilistic reasoning as a special case of probability updating. Methods that have been proposed for probability updating, therefore, also are candidates to solve inductive probabilistic inference problems.

For updating a unique prior distribution on categorical information, no viable alternative exists to *conditioning*: the posterior distribution is the prior conditioned on the stated facts [2]. Note that conditioning, seen as a rule for inductive reasoning, rather than probability updating, is just direct inference again.

As our examples already have shown, this basic updating/inductive reasoning problem can be generalized in two ways: first, the new information may come

[2] Lewis (1976) proposes *imaging* as an alternative to conditioning, but imaging requires a similarity measure on the states of the probability space, which usually cannot be assumed as given.

in the form of probabilistic constraints as in (7), not in the form of categorical statements; second, the prior (or statistical) information may be incomplete, and only specify a set of possible distributions as in (6), not a unique distribution. The problem of updating such partially defined beliefs has received considerable attention, e.g. (Dempster 1967, Shafer 1976, Walley 1991, Dubois & Prade 1997, Gilboa & Schmeidler 1993, Moral & Wilson 1995, Grove & Halpern 1998). The simplest approach is to apply an updating rule for unique priors to each of the distributions that satisfy the prior constraints, and to infer as partial posterior beliefs only probability assignments that are valid for all updated possible priors. Inferences obtained in this manner can be quite weak, and other principles have been explored where updating is performed only on a subset of possible priors that are in some sense maximally consistent with the new information (Gilboa & Schmeidler 1993, Dubois & Prade 1997). These methods are more appropriate for belief updating than for inductive probabilistic reasoning in our sense, because they amount to a combination of prior and new information on a more or less symmetric basis. As discussed above, this is not appropriate in our setting, where the new single case information is not supposed to have any impact on the statistical background knowledge. Our treatment of incompletely specified priors, therefore, follows the first approach of taking every possible prior (statistical distribution) into account. See section 4.1 for additional comments on this issue.

The main problem we address in the present paper is how to deal with new (single case) information in the form of general probability constraints. For this various rules with different scope of application have previously been explored. In the case where the new constraints prescribe the probability values $p_1, \ldots, p_k$ of pairwise disjoint alternatives $A_1, \ldots, A_k$, *Jeffrey's rule* (Jeffrey 1965) is a straightforward generalization of conditioning: it says that the posterior should be the sum of the conditional distributions given the A_i, weighted with the prescribed values p_i. Applying Jeffrey's rule to (1) and (7), for instance, we would obtain $0.6 \cdot 0.027 + 0.4 \cdot r$ as the probability for Jones getting into an accident, where r is the (unspecified) statistical probability of getting into an accident among drivers who do less than 10,000 or more than 20,000 miles.

When the constraints on the posterior are of a more general form than permitted by Jeffrey's rule, there no longer exist updating rules with a similarly intuitive appeal. However, a number of results indicate that *cross-entropy minimization* is the most appropriate general method for probability updating, or inductive probabilistic inference (Shore & Johnson 1980, Paris & Vencovská 1992, Jaeger 1995*b*). Cross-entropy can be interpreted as a measure for the similarity of two probability distributions (originally in an information theoretic sense (Kullback & Leibler 1951)). Cross-entropy minimization, therefore, is a rule according to which the posterior (or the subjective) distribution is chosen so as to make it as similar as possible within the given constraints to the prior (resp. the statistical) distribution.

Inductive probabilistic reasoning as we have explained it so far clearly is a topic with its roots in epistemology and the philosophy of science rather than in computer science. However, it also is a topic of substantial interest in all areas of artificial intelligence where one is concerned with reasoning and decision making under uncertainty.

Our introductory example is a first case in point. The inference patterns described in this example could be part of a probabilistic expert system employed by an insurance company to determine the rate of a liability insurance for a specific customer.

As a second example, consider the case of an autonomous agent that has to decide on its actions based on general rules it has been programmed with, and observations it makes. To make things graphic, consider an unmanned spacecraft trying to land on some distant planet. The spacecraft has been instructed to choose one of two possible landing sites: site A is a region with a fairly smooth surface, but located in an area subject to occasional severe storms; site B lies in a more rugged but atmospherically quiet area. According to the statistical information the spacecraft has been equipped with, the probabilities of making a safe landing are 0.95 at site A when there is no storm, 0.6 at site A under stormy conditions, and 0.8 at site B. In order to find the best strategy for making a safe landing, the spacecraft first orbits the planet once to take some meteorological measurements over site A. Shortly after passing over A it has to decide whether to stay on course to orbit the planet once more, and then land at A (20 hours later, say), or to change its course to initiate landing at B. To estimate the probabilities of making a safe landing following either strategy, thus the probability of stormy conditions at A in 20 hours time has to be evaluated. A likely method to obtain such a probability estimate is to feed the measurements made into a program that simulates the weather development over 20 hours, to run this simulation, say, one hundred times, each time adding some random perturbation to the initial data and/or the simulation, and to take the fraction q of cases in which the simulation at the end indicated stormy conditions at A as the required probability. Using Jeffrey's rule, then $0.6q + 0.95(1 - q)$ is the estimate for the probability of a safe landing at A.

This example illustrates why conditioning as the sole instrument of probabilistic inference is not enough: there is no way that the spacecraft could have been equipped with adequate statistical data that would allow it to compute the probability of storm at A in 20 hours time simply by conditioning the statistical data on its evidence, consisting of several megabytes of meteorological measurements. Thus, even a perfectly rational, automated agent, operating on the basis of a well-defined finite body of input data cannot always infer subjective probabilities by conditioning statistical probabilities, but will sometimes have to engage in more flexible forms of inductive probabilistic reasoning. [3]

[3] Jeffrey (1965) argues the same point for human reasoners with his "observation by candlelight"-example. That argument, however, is not directly transferable to an autonomous agent whose evidence – at least in principle – is always expressible by a single, well-defined, proposition

1.2 Aims and Scope

To make inductive probabilistic reasoning available for AI applications, two things have to be accomplished: first, a formal rule for this kind of probabilistic inference has to be found. Second, a formal representation language has to be developed that allows us to encode the kind of probabilistic statements we want to reason with, and on which inference rules for inductive probabilistic reasoning can be defined.

In this paper we will focus on the second of these problems, basically taking it for granted that cross-entropy minimization is the appropriate formal rule for inductive probabilistic reasoning (see section 3.1 for a brief justification). The representation language that we will develop is first-order predicate logic with additional constructs for the representation of statistical and subjective probability statements. To encode both deductive and inductive inferences on this language, it will be equipped with two different entailment relations: a relation $\models$ that describes valid probabilistic inferences, and a relation $|\!\!\approx$ that describes inductive probabilistic inferences obtained by cross-entropy minimization. For example, the representation language will be rich enough to encode all the example statements (1)-(8) in formal sentences $\phi_1, \ldots, \phi_8$.

If, furthermore, ψ_0 is a sentence that says that with probability 0.4 Jones drives less than 10000 or more than 20000 miles annually, then we will obtain in our logic

$$\phi_7 \models \psi_0,$$

because ψ_0 follows from ϕ_7 by the laws of probability theory. If, on the other hand, ψ_1 says that with probability at least 0.0162 Jones will be involved in an accident, then ψ_1 does not strictly follow from our premises, i.e.

$$\phi_1 \wedge \phi_7 \not\models \psi_1.$$

However, for the inductive entailment relation we will obtain

$$\phi_1 \wedge \phi_7 \mathrel{|\!\!\approx} \psi_1.$$

Our probabilistic first-order logic with the two entailment relations $\models$ and $|\!\!\approx$ will provide a principled formalization of inductive probabilistic reasoning in an expressive logical framework. The next problem, then, is to define inference methods for this logic. It is well-known that for probabilistic logics of the kind we consider here no complete deduction calculi exist when probabilities are required to be real numbers (Abadi & J.Y.Halpern 1994), but that completeness results can be obtained when probability values from more general algebraic structures are permitted (Bacchus 1990*a*). We will follow the approach of generalized probabilities and permit probabilities to take values in *logarithmic real-closed fields (lrc-fields)*, which provide a very good approximation to the real numbers. With the lrc-field based semantics we obtain a completeness result for our logic. It should be emphasized that with this approach we do not abandon real-valued probabilities: real numbers being an example for an lrc-field, they are, of course, not excluded by our generalized semantics. Moreover, a completeness result for lrc-field valued probabilities can also be read as a characterization of the degree of incompleteness of our deductive system for real-valued probabilities: the only inferences for real-valued probabilities that we are not able to make are those that are not valid in all other lrc-fields. By complementing the completeness result for lrc-field valued probabilities with results showing that core properties of real-valued probabilities are actually shared by all lrc-field valued probabilities, we obtain a strong and precise characterization of how powerful our deductive system is for real-valued probabilities.

The main part of this paper (sections 2 and 3) contains the definition of our logic $\mathscr{L}_{ip}$ consisting of a probabilistic representation language L_p, a strict entailment relation $\models$ (both defined in section 2), and an inductive entailment relation $\mid\!\approx$ (defined in section 3). The basic design and many of the properties of the logic $\mathscr{L}_{ip}$ do not rely on our use of probability values from logarithmic real-closed fields, so that sections 2 and 3 can also be read ignoring the issue of generalized probability values, and thinking of real-valued probabilities throughout. Only the key properties of $\mathscr{L}_{ip}$ expressed in corollary 2.11 and theorem 2.12 are not valid for real-valued probabilities.

To analyze in detail the implications of using lrc-fields we derive a number of results on cross-entropy and cross-entropy minimization in logarithmic real-closed fields. The basic technical results here have been collected in appendix A. These results are used in section 3 to show that many important inference patterns for inductive probabilistic reasoning are supported in $\mathscr{L}_{ip}$. The results of appendix A also are of some independent mathematical interest, as they constitute an alternative derivation of basic properties of cross-entropy minimization in (real-valued) finite probability spaces only from elementary algebraic properties of the logarithmic function. Previous derivations of these properties required more powerful analytic methods (Kullback 1959, Shore & Johnson 1980).

This paper is largely based on the author's PhD thesis (Jaeger 1995*a*). A very preliminary exposition of the logic $\mathscr{L}_{ip}$ was was given in (Jaeger 1994*a*). A statistical derivation of cross-entropy minimization as the formal model for inductive probabilistic reasoning was given in (Jaeger 1995*b*).

1.3 Previous Work

Clearly, the work here presented is intimately related to a sizable body of previous work on combining logic and probability, and on the principles of (probabilistic) inductive inference.

Boole (1854) must probably be credited for being the first to combine logic and probability. He saw events to which probabilities are attached as formulas in a (propositional) logic, and devised probabilistic inference techniques that were based both on logical manipulations of the formulas and algebraic techniques for solving systems of (linear) equations (see (Hailperin 1976) for a modern exposition of Boole's work).

The work of Carnap (1950, 1952) is of great interest in our context in more than one respect: Carnap was among the first to acknowledge the existence of two legitimate concepts of probability, (in Carnap's terminology) expressing degrees of confirmation and relative frequencies, respectively. The main focus in Carnap's work is on probability as degree of confirmation, which he considers to be defined on logical formulas. His main objective is to find a canonical probability distribution $\mathfrak{c}$ on the algebra of (first-order) formulas, which would allow to compute the degree of confirmation $\mathfrak{c}(h/e)$ of some hypothesis h, given evidence e in a mechanical way, i.e. from the syntactic structure of h and e alone. Such a confirmation function $\mathfrak{c}$ would then be seen as a normative rule for inductive reasoning. While eventually abandoning the hope to find such a unique confirmation function (Carnap 1952), Carnap (1950) proves that for a general class of candidate functions $\mathfrak{c}$ a form of the direct inference principle can be derived: if e is a proposition that says that the relative frequency of some property M in a population of n objects is r, and h is the proposition that one particular of these n objects has property M, then $\mathfrak{c}(h/e) = r$.

Carnap's work was very influential, and many subsequent works on probability and logic (Gaifman 1964, Scott & Krauss 1966, Fenstad 1967, Gaifman & Snir 1982) were more or less directly spawned by (Carnap 1950). They are, however, more concerned with purely logical and mathematical questions arising out of the study of probabilistic interpretations for logical language, than with the foundations of probabilistic and inductive reasoning.

In none of the works mentioned so far were probabilistic statements integrated into the logical language under consideration. Only on the semantic level were probabilities assigned to (non-probabilistic) formulas. This changes with Kyburg (1974), who, like Carnap, aims to explain the meaning of probability by formalizing it in a logical framework. In doing so, he develops within the framework of first-order logic special syntactic constructs for statistical statements. These statistical statements, in conjunction with a body of categorical knowledge, then are used to define subjective probabilities via direct inference.

Keisler (1985) and Hoover (1978) developed first-order and infinitary logics in which the standard quantifiers $\forall x$ and $\exists x$ are replaced by a probability quantifier $Px \geq r$, standing for "for x with probability at least r". The primary motivation behind this work was to apply new advances in infinitary logics to probability theory.

In AI, interest in probabilistic logic started with Nilsson's (1986) paper, which, in many aspects, was a modern reinvention of (Boole 1854) (see (Hailperin 1996) for an extensive discussion).

Halpern's (1990) and Bacchus's (1990*b*, 1990*a*) seminal works introduced probabilistic extensions of first-order logic for the representation of both statistical and subjective probabilities within the formal language. The larger part of Halpern's and Bacchus's work is concerned with coding strict probabilistic inferences in their logics. A first approach towards using the underlying probabilistic logics also for inductive probabilistic reasoning is contained in (Bacchus 1990*b*), where an axiom schema for direct inference is presented. Much more general patterns of inductive (or default) inferences are modeled by the *random worlds* method by Bacchus, Grove, Halpern, and Koller (Bacchus et al. (1992, 1997), Grove et al. (1992*a*, 1992*b*)). By an approach very similar to Carnap's definition of the confirmation function $\mathfrak{c}$, in this method a degree of belief $\Pr(\phi|\psi)$ in ϕ given the knowledge ψ is defined. Here ϕ and ψ now are formulas in the statistical probabilistic languages of Halpern and Bacchus. As ψ, thus, cannot encode prior constraints on the subjective probabilities (or degrees of belief), the reasoning patterns supported by this method are quite different from what we have called inductive probabilistic reasoning in section 1.1, and what forms the subject of the current paper. A more detailed discussion of the random worlds method and its relation to our framework is deferred to section 4.1.

2 The Logic of Strict Inference

2.1 Outline

In this section we introduce the logic $\mathscr{L}_p = (L_p, \models)$ consisting of a language L_p for the representation of statistical and subjective probabilities, and an entailment relation $\models$ capturing inferences that are validated by probability calculus. Thus, the nature of the logic $\mathscr{L}_p$ will be very similar to that of the logics of Halpern (1990) and Bacchus (1990*b*), and we will follow in our presentation of $\mathscr{L}_p$ these previously defined formalisms as far as possible.

The main difference between our logic $\mathscr{L}_p$ and the logics of Halpern and Bacchus lies in the definition of terms expressing subjective probabilities. Here our approach is guided by the goal to later extend the logic $\mathscr{L}_p$ to a logic $\mathscr{L}_{ip} = (L_p, \models, \mid\!\approx)$ with an additional entailment relation $\mid\!\approx$ for inductive probabilistic inferences. This inductive entailment relation will be obtained by implementing cross-entropy minimization between the statistical and subjective probability distribution in the semantic structures for the language. As we can only speak of the cross-entropy of two probability distributions that are defined on the same probability space, we cannot follow Bacchus and Halpern in interpreting statistical and subjective probability terms by probability distributions over the domains of semantical structures, and distributions over sets of semantic structures, respectively. Instead, we choose to interpret both statistical and subjective probability terms over the domain of semantic structures. To make this feasible for subjective probability terms, we have to impose a certain restriction on their formulation: it will be required that subjective probability terms always refer to some specific objects or events about which there is some uncertainty. In our introductory example, for instance, all the uncertainty expressed in the subjective probability statements was attached to the object "Jones" about whose exact properties we have incomplete information. In a somewhat more complicated example, a subjective probability statement may be about the probability that in an accident "crash010899Madison/5th", involving drivers "Jones" and "Mitchell", driver "Jones" was to be blamed for the accident. This statement, then, would express uncertainty about the exact relations between the elements of the tuple (crash010899Madison/5th,Jones,Mitchell) of objects and events.

Considering only subjective probability expressions that fit this pattern allows us to interpret them by probability distributions over the domain of a semantic structure: we interpret the concrete objects and events appearing in the subjective probability expression as randomly drawn elements of the domain. This approach stands in the tradition of frequentist interpretations of subjective probabilities (Reichenbach 1949, Carnap 1950). For the denotation of such random domain elements we will use a special type of symbols, called *event symbols*, that are used syntactically like constants, but are interpreted by probability measures.

Another point where we will deviate from the previous approaches by Halpern and Bacchus is in the structure of the probability measures appearing as part of the semantic structures. In (Halpern 1990) and (Bacchus 1990*b*) these measures were assumed to come from the very restricted class of real-discrete measures (cf. example 2.7 below). Halpern (1990) states that this restriction is not essential and briefly outlines a more general approach, perhaps somewhat understating the technical difficulties arising in these approaches (as exemplified by our theorem 2.8 below). In Bacchus (1990*a*) a more general concept of probability distributions is used, allowing arbitrary finitely additive field-valued probabilities. We will use a closely related approach, requiring probabilities to take values in *logarithmic real-closed fields* (definition 2.1 below).

2.2 Syntax

The syntax of our logic is that of first-order predicate logic with three extensions: first, the language of logarithmic, ordered fields is integrated as a fixed component into the language; second, a term-forming construction (taken directly from Bacchus (1990*b*)) is introduced that allows us to build terms denoting statistical probabilities; and third, a term-forming construction is introduced for building terms denoting subjective probabilities.

We use two sets of variables in the language: *domain variables* ranging over the elements of the domain of discourse, and *field variables* ranging over numbers, especially probability values. The vocabulary

$$\mathrm{S}_{\mathrm{LOF}} = \{0, 1, +, \cdot, \leq, \mathit{Log}\}$$

of ordered fields with a logarithmic function is considered to belong to the logical symbols of the language. The non-logical symbols consist of a set $S = \{\mathtt{R}, \mathtt{Q}, \ldots, \mathtt{f}, \mathtt{g}, \ldots, \mathtt{c}, \mathtt{d}, \ldots\}$ of relation, function, and constant symbols, as in first-order logic, and a tuple $\mathbf{e} = (\mathrm{e}_1, \ldots, \mathrm{e}_N)$ of *event symbols*.

The language $L_p(S, \mathbf{e})$ now is defined by the following rules. Since in part (f) of the formation rule for field terms a condition on the free variables of a formula is required, we have to define simultaneously with the construction of terms and formulas the set of free variables they contain. Except for the nonstandard syntactic constructions we omit these obvious declarations.

A *domain-term* is constructed from domain-variables $v_0, v_1, \ldots$, constant and function symbols from S according to the syntax rules of first-order logic.

Atomic domain formulas are formulas of the form

$$\mathtt{R}\,\mathrm{t}_1 \ldots \mathrm{t}_k \quad \text{or} \quad \mathrm{t}_1 = \mathrm{t}_2,$$

where $\mathtt{R}$ is a k-ary relation symbol from S, and the t_i are domain-terms.

Boolean operations: If ϕ and ψ are formulas, then so are $(\phi \wedge \psi)$ and $\neg\phi$.

Quantification: If ϕ is a formula and v (x) is a domain-variable (field-variable), then $\exists v\phi$ ($\exists x\phi$) is a formula.

Field-terms:

(a) Every field-variable $x_0, x_1, \ldots$ is a field-term.

(b) 0 and 1 are field-terms

(c) If t_1 and t_2 are field-terms, then so are $(\mathrm{t}_1 \cdot \mathrm{t}_2)$ and $(\mathrm{t}_1 + \mathrm{t}_2)$.

(d) If t is a field term, then so is *Log*(t).

(e) If ϕ is a formula, and $\boldsymbol{w}$ a tuple of domain variables, then

$$[\phi]_{\boldsymbol{w}}$$

is a field-term. The free variables of $[\phi]_{\boldsymbol{w}}$ are the free variables of ϕ not appearing in $\boldsymbol{w}$. A field term of this form is called a *statistical probability term.*

(f) If $\phi(\boldsymbol{v})$ is a formula whose free variables are among the domain variables $\boldsymbol{v}$, ϕ does not contain any terms of the form prob(...), and if $\boldsymbol{v}/\mathbf{e}$ is an assignment that maps every $v \in \boldsymbol{v}$ to some $\mathrm{e} \in \mathbf{e}$, then

$$\mathrm{prob}(\phi[\boldsymbol{v}/\mathbf{e}])$$

is a field-term (without free variables). A field term of this form is called a *subjective probability term.*

Atomic field formulas: If $\mathrm{t}_1, \mathrm{t}_2$ are field-terms, then $\mathrm{t}_1 \leq \mathrm{t}_2$ is an atomic field formula.

Rule (f) for field terms essentially says that event symbols $\mathrm{e}_1, \ldots, \mathrm{e}_N$ are used syntactically like constant symbols, but are restricted to only appear within the scope of a prob()-operator. Moreover, subjective probability terms may not be nested or contain free variables. These are fairly serious limitation that are not essential for the definition of $\mathscr{L}_p$, but will be crucially important for the definition of $\models\!\approx$ in $\mathscr{L}_{ip}$.

We may freely use as definable abbreviations (in)equalities like $t_1 > t_2$, $t_1 = t_2$, $t_1 \geq t_2$, and conditional probability expressions like $[\phi \mid \psi]_{\boldsymbol{w}}$ or $\mathrm{prob}(\phi[\mathbf{e}] \mid \psi[\mathbf{e}])$. These conditional probability expressions are interpreted by the quotients $[\phi \wedge \psi]_{\boldsymbol{w}}/[\psi]_{\boldsymbol{w}}$, respectively $\mathrm{prob}(\phi[\mathbf{e}] \wedge \psi[\mathbf{e}])/\mathrm{prob}(\psi[\mathbf{e}])$, provided the interpretations of $[\psi]_{\boldsymbol{w}}$, respectively $\mathrm{prob}(\psi[\mathbf{e}])$, are positive. Several conventions may be employed to interpret conditional probability terms when the conditioning expressions are assigned probability zero. We will not explore this issue here and refer the reader to (Bacchus 1990*b*), (Halpern 1990), and (Jaeger 1995*a*) for alternative proposals.

To illustrate the use of the language L_p, we encode some of the example sentences of section 1.1. We use a vocabulary that contains two unary predicate symbols D and M that partition the domain into elements of the sorts driver and mileage, respectively. Another unary predicate symbol IIA stands for "involved in accident", and a unary function am maps drivers to their annual mileage. Also we use constants 10,15,... for specific mileages (in thousands), and a binary order relation $\preceq$ on mileages (this relation $\preceq$ defined on the domain is to be distinguished from the relation $\leq$ defined on probability values). Finally, there is a single event symbol jones. Statement (1) can now be formalized as

$$\phi_1 :\equiv \; [\mathtt{IIA}(d) \mid \mathtt{D}(d) \wedge 10 \preceq \mathtt{am}(d) \preceq 20]_d = 0.027. \tag{9}$$

Statement (3) becomes

$$\phi_3 :\equiv \mathrm{prob}(\mathtt{IIA}(jones)) = 0.027. \tag{10}$$

2.3 Semantics

Key components of the semantic structures that we will use to interpret L_p are finitely additive probability measures with values in logarithmic real-closed fields. We briefly review the concepts we require.

Definition 2.1 An $\mathrm{S}_{\mathrm{LOF}}$-structure $\mathfrak{F}$=$(\mathbb{F}, 0, 1, +, \cdot, \leq, \mathit{Log})$ over a domain $\mathbb{F}$ is a *logarithmic real closed field* (lrc-field for short), if it satisfies the axioms LRCF consisting of

(i) The axioms of ordered fields.

(ii) An axiom for the existence of square roots:

$$\forall x \exists y (0 \leq x \rightarrow y^2 = x).$$

(iii) A schema demanding that every polynomial of uneven degree has a root:

$$\forall y_0 \dots y_{n-1} \exists x (y_0 + y_1 \cdot x + \dots + y_{n-1} \cdot x^{n-1} + x^n = 0). \quad n = 1, 3, 5, \dots$$

(iv) $\forall x, y > 0 \;\; \mathit{Log}(x \cdot y) = \mathit{Log}(x) + \mathit{Log}(y)$

(v) $\forall x > 0 \;\; x \neq 1 \rightarrow \mathit{Log}(x) < x - 1$

(viii) The approximation schema

$$\forall x \in (0,1] \;\; q_n(x) \leq \mathit{Log}(x) \leq p_n(x) \qquad (n = 1, 2, \dots)$$

where

$$\begin{aligned} q_n(x) &:\equiv (x-1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} - \dots + (-1)^{n-1}\frac{(x-1)^n}{n} + (-1)^n \frac{(x-1)^{n+1}}{x} \\ p_n(x) &:\equiv (x-1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} - \dots + (-1)^{n-1}\frac{(x-1)^n}{n}, \end{aligned}$$

A structure over the vocabulary $\mathrm{S_{OF}} := \{+, \cdot, \leq, 0, 1\}$ that satisfies the axioms RCF consisting of (i)-(iii) alone is called a real-closed field. By classic results in model theory, RCF is a complete axiomatization of the $\mathrm{S_{OF}}$-theory of the real numbers. In other words, every first-order $\mathrm{S_{OF}}$-sentence ϕ that is true in $\mathbb{R}$ also is true in every other real-closed field (see (Rabin 1977) for an overview). To what extent similar results hold for logarithmic real closed fields is a long-standing open problem in model theory (there studied w.r.t. (real-closed) fields augmented by an exponential, rather than a logarithmic, function; see e.g. (Dahn & Wolter 1983)).

Definition 2.2 Let M be a set. An *algebra* over M is a collection $\mathfrak{A}$ of subsets of M that contains M, and is closed under complementation and finite unions. If M is also closed under countable unions, it is called a *σ-algebra.* If $\mathfrak{A}$ is an algebra on M, and $\mathfrak{A}'$ an algebra on M', then the *product algebra* $\mathfrak{A} \times \mathfrak{A}'$ is the algebra on $M \times M'$ generated by the sets $A \times A'$ ($A \in \mathfrak{A}$, $A' \in \mathfrak{A}'$).

Definition 2.3 Let $\mathfrak{A}$ be an algebra over M, $\mathfrak{F}$ an lrc-field. Let $\mathbb{F}^+ := \{x \in \mathbb{F} \mid 0 \leq x\}$. A function

$$P : \mathfrak{A} \to \mathbb{F}^+$$

is an $\mathbb{F}$-*probability measure* iff $P(\emptyset) = 0$, $P(M) = 1$, and $P(A \cup B) = P(A) + P(B)$ for all $A, B \in \mathfrak{A}$ with $A \cap B = \emptyset$. The elements of $\mathfrak{A}$ also are called the *measurable* sets. The set of all probability measures with values in $\mathbb{F}$ on the algebra $\mathfrak{A}$ is denoted by

$$\Delta_{\mathbb{F}}\mathfrak{A}.$$

Thus, even when the underlying algebra is a σ-algebra, we do not require σ-additivity, because this would usually make no sense in arbitrary lrc-fields, where infinite sums of non-negative numbers need not be defined. If $\mathfrak{A}$ is a finite algebra with n atoms, then $\Delta_{\mathbb{F}}\mathfrak{A}$ can be identified with

$$\Delta^n_{\mathbb{F}} := \{(x_1, \ldots, x_n) \in \mathbb{F}^n \mid \ x_i \geq 0, \ \sum_i x_i = 1\}.$$

If $\mathfrak{A}'$ is a subalgebra of $\mathfrak{A}$, and $P \in \Delta_{\mathbb{F}}\mathfrak{A}$, then $P \restriction \mathfrak{A}'$ denotes the restriction of P to $\mathfrak{A}'$, i.e. a member of $\Delta_{\mathbb{F}}\mathfrak{A}'$. By abuse of notation we also use $P \restriction \mathfrak{A}'$ to denote the marginal distribution on $\mathfrak{A}'$ when $\mathfrak{A}'$ is a factor, rather than a subalgebra, of $\mathfrak{A}$, i.e. $\mathfrak{A} = \mathfrak{A}' \times \mathfrak{A}''$ for some $\mathfrak{A}''$.

Semantic structures for the interpretation of $L_p(S, \mathbf{e})$ are based on standard model theoretic structures for the vocabulary S, augmented by probability measures for the interpretation of probability terms.

The basic form of a probabilistic structure will be

$$\mathfrak{M} = (M, I, \mathfrak{F}, (\mathfrak{A}_n, P_n)_{n \in \mathbb{N}}, Q_{\mathbf{e}})$$

where (M, I) is a standard S-structure consisting of domain M and interpretation function I for S, $\mathfrak{F}$ is a logarithmic real closed field, the $(\mathfrak{A}_n, P_n)$ are probability measure algebras on M^n, and $Q_{\mathbf{e}}$ is a probability measure on $\mathfrak{A}_{|\mathbf{e}|}$ (we use $|\mathbf{e}|$, $|\boldsymbol{v}|$, etc., to denote the number of elements in a tuple of event symbols $\mathbf{e}$, variables $\boldsymbol{v}$, etc.).

Statistical probability terms $[\phi]_{\boldsymbol{w}}$ will be interpreted by $P_{|\boldsymbol{w}|}(A)$ where A is the set defined by ϕ in $M^{|\boldsymbol{w}|}$. The measure P_n, thus, is intended to represent the distribution of a sample of n independent draws from the domain, identically distributed according to P_1 (an "iid sample of size n"). In the case of real-valued σ-additive measures this would usually be achieved by defining P_n to be the n-fold product of P_1, defined on the product σ-algebra $\mathfrak{A}_1 \times \ldots \times \mathfrak{A}_1$ (n factors). A corresponding approach turns out to be infeasible in our context, because the product algebra $\mathfrak{A}_1 \times \ldots \times \mathfrak{A}_1$ usually will not be fine-grained enough to give semantics to all statistical probability terms $[\phi]_{\boldsymbol{w}}$. In order to ensure that the sequence $(\mathfrak{A}_1, P_1), (\mathfrak{A}_2, P_2), \ldots$, nevertheless, behaves in several essential aspects like a sequence of product algebras and product measures, we explicitly impose three *coherence conditions*: *homogeneity*, the *product property*, and the *Fubini property.* These are essentially the same conditions as can be found in (Hoover 1978), there summarily called Fubini property. Bacchus (1990*a*) requires homogeneity and the product property only.

Homogeneity: For all n, $A \in \mathfrak{A}_n$ and permutations π of $\{1, \ldots, n\}$:

$$\pi(A) := \{\pi\boldsymbol{a} \mid \boldsymbol{a} \in A\} \in \mathfrak{A}_n, \quad \text{and} \quad P_n(\pi(A)) = P_n(A).$$

Homogeneity expresses the permutation invariance of iid samples: if we sample two drivers from our example domain, for instance, then the probability that the first one drives a Toyota, and the second one a Ford is the same as the probability that the first one drives a Ford, and the second one a Toyota.

Product property: For all $k, l \in \mathbb{N}$: $A \in \mathfrak{A}_k$ and $B \in \mathfrak{A}_l$ implies $A \times B \in \mathfrak{A}_{k+l}$, and $P_{k+l}(A \times B) = P_k(A) \cdot P_l(B)$.

The product property expresses independence of samples. For an example let $k = l = 1$, A comprise the set of Toyota drivers, and B comprise the set of Ford drivers. Then $P_1(A)$ ($P_1(B)$) is the probability of sampling a Toyota (Ford) driver in a single draw. $P_2(A \times B)$ is the probability of first drawing a Toyota driver, then a Ford driver, in a two-element sample. When sampling is iid, $P_2(A \times B)$ must be equal to $P_1(A)P_1(B)$.

For the formulation of the last coherence condition we first introduce some notation for sections of sets: Let $I \subset \{1, \ldots, n\}$ with $I \neq \emptyset$ and $I' := \{1, \ldots, n\} \backslash I$. Let $A \subseteq M^n$ and $\boldsymbol{a} \in M^I$. Then the *section* of A in the coordinates I along $\boldsymbol{a}$ is defined as

$$\sigma^I_{\boldsymbol{a}}(A) := \{\boldsymbol{b} \in M^{I'} \mid (\boldsymbol{a}, \boldsymbol{b}) \in A\}.$$

Fubini property: For all $n \in \mathbf{N}$, $I \subset \{1, \ldots, n\}$ with $1 \leq |I| =: k$, $A \in \mathfrak{A}_n$, and $\boldsymbol{a} \in M^I$:

$$\sigma^I_{\boldsymbol{a}}(A) \in \mathfrak{A}_{n-k}, \tag{11}$$

for all $r \in [0, 1]$:

$$A_{I,\geq r} := \{\boldsymbol{a} \in M^I \mid P_{n-k}(\sigma^I_{\boldsymbol{a}}(A)) \geq r\} \in \mathfrak{A}_k, \tag{12}$$

and

$$P_n(A) \geq r P_k(A_{I,\geq r}). \tag{13}$$

Furthermore, we require (13) to hold with strict inequality for the set $A_{I,>r}$ defined by replacing $\geq$ by $>$ in (12).

The Fubini property expresses a fundamental "commensurability" property of product measures in different dimensions. For standard σ-additive measures it plays a vital role in the theory of integration. It is best illustrated by a geometric example: obviously, if a geometric figure A in the plane contains a rectangle with sides of lengths s and r, then the area of A must be at least $r \cdot s$. This is essentially the defining property of area as the product measure of one-dimensional lengths. Furthermore, the lower bound $r \cdot s$ also holds when A only contains a "distorted rectangle" of dimensions $r \times s$, as illustrated in figure 1. The Fubini property establishes the lower bound of $r \cdot s$ for the measure of A from a condition that further generalizes the property of containing a "distorted rectangle".

We are now ready to define our semantic structures.

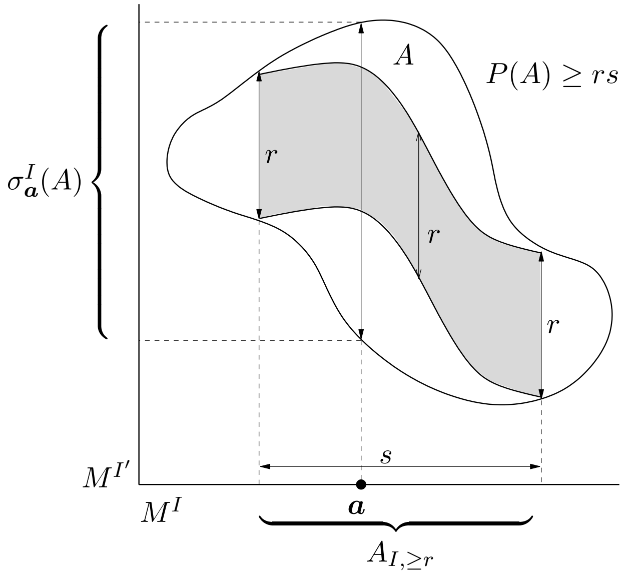

Figure 1: The Fubini property

Definition 2.4 Let S be a vocabulary, $\mathbf{e}$ a tuple of event symbols. A *probabilistic structure* for $(S, \mathbf{e})$ is a tuple

$$\mathfrak{M} = (M, I, \mathfrak{F}, (\mathfrak{A}_n, P_n)_{n \in \mathbb{N}}, Q_{\mathbf{e}})$$

where M is a set (the domain), I is an interpretation function for S over M, $\mathfrak{F}$ is a lrc-field, $(\mathfrak{A}_n, P_n)$ is a measure algebra on M^n $(n \in \mathbb{N})$, such that the sequence $(\mathfrak{A}_n, P_n)_{n \in \mathbb{N}}$ satisfies homogeneity, the product property, and the Fubini property, and $Q_{\mathbf{e}}$ is a probability measure on $\mathfrak{A}_{|\mathbf{e}|}$.

Now let a probabilistic structure $\mathfrak{M}$ for $(S, \mathbf{e})$ be given, let γ be a variable assignment that maps domain-variables into M and field-variables into $\mathbb{F}$. The notation $\gamma[\boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r}]$ is used for the variable assignment that maps $\boldsymbol{v}$ to $\boldsymbol{a}$, $\boldsymbol{x}$ to $\boldsymbol{r}$, and for all other variables is the same as γ.

We now need to define the satisfaction relation between $(\mathfrak{M}, \gamma)$ and L_p-formulas. Due to the possible non-measurability of L_p-definable sets, this relation may only be partial. In detail, we define a partial interpretation that maps an $(S, \mathbf{e})$-term t to its interpretation $(\mathfrak{M}, \gamma)(\mathrm{t})$ in M (if it is a domain term), or in $\mathbb{F}$ (if it is a field term). In parallel, a relation $\models$ is defined between $(\mathfrak{M}, \gamma)$ and $L_p(S, \mathbf{e})$-formulas ϕ. This relation,too, may be only partial in the sense that it is possible that neither $(\mathfrak{M}, \gamma) \models \phi$, nor $(\mathfrak{M}, \gamma) \models \neg\phi$.

Domain-terms: For a domain-term t, the interpretation $(\mathfrak{M}, \gamma)(\mathrm{t})$ is defined just as in first-order logic. Note that t cannot contain any field-terms as subterms.

Atomic domain formulas: If ϕ is an atomic domain formula then the relation $(\mathfrak{M}, \gamma) \models \phi$ is defined as in first-order logic.

Boolean operations: The definition of $(\mathfrak{M}, \gamma) \models \phi$ for $\phi = \psi \vee \chi$ and $\phi = \neg\psi$ is as usual, provided that $\models$ is defined between $(\mathfrak{M}, \gamma)$ and the subformulas ψ, χ. Otherwise $\models$ is undefined between $(\mathfrak{M}, \gamma)$ and ϕ.

Quantification: Let $\phi(\boldsymbol{v}, \boldsymbol{x}) \equiv \exists w \psi(\boldsymbol{v}, w, \boldsymbol{x})$. Then

$$(\mathfrak{M}, \gamma) \models \phi(\boldsymbol{v}, \boldsymbol{x}) \text{ iff } \exists a \in M \quad (\mathfrak{M}, \gamma[w/a]) \models \psi(\boldsymbol{v}, w, \boldsymbol{x}).$$

Similarly for quantification over field variables and universal quantification.

Field-terms: Let t be a field-term.

(a) $\mathrm{t} \equiv x$. Then $(\mathfrak{M}, \gamma)(\mathrm{t}) = \gamma(x)$.

(b) $\mathrm{t} \equiv 0$. Then $(\mathfrak{M}, \gamma)(\mathrm{t}) = 0$. Similarly for $\mathrm{t} \equiv 1$.

(c) $\mathrm{t} \equiv \mathrm{t}_1 + \mathrm{t}_2$. Then $(\mathfrak{M}, \gamma)(\mathrm{t}) = (\mathfrak{M}, \gamma)(\mathrm{t}_1) + (\mathfrak{M}, \gamma)(\mathrm{t}_2)$ if $(\mathfrak{M}, \gamma)(\mathrm{t}_1)$ and $(\mathfrak{M}, \gamma)(\mathrm{t}_2)$ are defined. $(\mathfrak{M}, \gamma)(\mathrm{t})$ is undefined otherwise. Similarly for $\mathrm{t} \equiv \mathrm{t}_1 \cdot \mathrm{t}_2$.

(d) $\mathrm{t} \equiv \mathit{Log}(\mathrm{t}')$. Then $(\mathfrak{M}, \gamma)(\mathrm{t}) = \mathit{Log}((\mathfrak{M}, \gamma)(\mathrm{t}'))$ if $(\mathfrak{M}, \gamma)(\mathrm{t}')$ is defined. $(\mathfrak{M}, \gamma)(\mathrm{t})$ is undefined otherwise.

(e) $\mathrm{t} \equiv [\phi(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{x})]_{\boldsymbol{w}}$. Then

$$(\mathfrak{M}, \gamma)(\mathrm{t}) = P_{|\boldsymbol{w}|}(\{\boldsymbol{a} \mid (\mathfrak{M}, \gamma[\boldsymbol{w}/\boldsymbol{a}]) \models \phi(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{x})\}),$$

if $\{\boldsymbol{a} \mid (\mathfrak{M}, \gamma[\boldsymbol{w}/\boldsymbol{a}]) \models \phi(\boldsymbol{u}, \boldsymbol{w}, \boldsymbol{x})\} \in \mathfrak{A}_{|\boldsymbol{w}|}$; $(\mathfrak{M}, \gamma)(\mathrm{t})$ is undefined otherwise.

(f) $\mathrm{t} \equiv \mathrm{prob}(\phi[\boldsymbol{v}/\mathbf{e}])$. Then

$$(\mathfrak{M}, \gamma)(\mathrm{t}) = Q_{\mathbf{e}}(\{\boldsymbol{a} \mid (\mathfrak{M}, \gamma[\boldsymbol{v}/\boldsymbol{a}]) \models \phi(\boldsymbol{v})\})$$

if $\{\boldsymbol{a} \mid (\mathfrak{M}, \gamma[\boldsymbol{v}/\boldsymbol{a}]) \models \phi(\boldsymbol{v})\} \in \mathfrak{A}_{|\mathbf{e}|}$; $(\mathfrak{M}, \gamma)(\mathrm{t})$ is undefined otherwise.

Atomic field formulas: Let $\phi \equiv \mathrm{t}_1 \leq \mathrm{t}_2$. Then $(\mathfrak{M}, \gamma) \models \phi$ iff $(\mathfrak{M}, \gamma)(\mathrm{t}_1)$ and $(\mathfrak{M}, \gamma)(\mathrm{t}_2)$ are defined, and $(\mathfrak{M}, \gamma)(\mathrm{t}_1) \leq (\mathfrak{M}, \gamma)(\mathrm{t}_2)$.

Definition 2.5 A probabilistic structure $\mathfrak{M}$ is *sufficient* if the relation $(\mathfrak{M}, \gamma) \models \phi$ is defined for all γ and all $\phi \in L_p$.

In other words, $\mathfrak{M}$ is sufficient if all L_p-definable sets are measurable. We define semantic entailment with respect to sufficient structures only:

Definition 2.6 For $\Phi \subseteq L_p$, $\psi \in L_p$ we write $\Phi \models \psi$ if for all sufficient probabilistic structures $\mathfrak{M}$: $(\mathfrak{M}, \gamma) \models \Phi$ implies $(\mathfrak{M}, \gamma) \models \psi$.

Because of the importance of definability, we introduce a somewhat more compact notation for sets defined by formulas: if ϕ is an $L_p(S, \mathbf{e})$-formula, $\mathfrak{M}$ a probabilistic structure, γ a variable assignment, and $\boldsymbol{v}$ a tuple of n distinct domain variables, then we write

$$(\mathfrak{M}, \gamma, \boldsymbol{v})(\phi) := \{\boldsymbol{a} \in M^n \mid (\mathfrak{M}, \gamma[\boldsymbol{v}/\boldsymbol{a}]) \models \phi\}. \tag{14}$$

Furthermore, when $\phi \equiv \phi(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{x})$, $\gamma(\boldsymbol{w}) = \boldsymbol{b}$, and $\gamma(\boldsymbol{x}) = \boldsymbol{r}$, then we also denote (14) by $(\mathfrak{M}, \boldsymbol{v})(\phi(\boldsymbol{v}, \boldsymbol{b}, \boldsymbol{r}))$.

It can be very difficult to verify sufficiency for a given structure $\mathfrak{M}$. In fact, the only class of examples of probabilistic structures for which sufficiency is easily proved is the following.

Example 2.7 Let S be a vocabulary, $\mathbf{e} = (e_1, \ldots, e_N)$ a tuple of event symbols. Let (M, I) be a standard S-structure; for $i \in \mathbb{N}$ let $a_i \in M, \boldsymbol{b}_i \in M^N, p_i, q_i \in \mathbb{R}$ with $\sum p_i = \sum q_i = 1$. Let $\mathfrak{A}_n = 2^{M^n}$ for all $n \in \mathbb{N}$, and define

$$P_n(A) = \sum_{(a_{i_1}, \ldots, a_{i_n}) \in A} p_{i_1} \cdot \ldots \cdot p_{i_n} \quad (A \subseteq M^n),$$

and

$$Q_{\mathbf{e}}(A) = \sum_{\boldsymbol{b}_i \in A} q_i \quad (A \subseteq M^N).$$

It is easy to see that $(\mathfrak{A}_n, P_n)_{n\in\mathbb{N}}$ satisfies the coherency conditions. Moreover, sufficiency is trivially satisfied, because every subset of M^n is measurable. We refer to structures of this form as *real-discrete structures.*

2.4 Probabilistic Reasoning in $\mathscr{L}_p$

The logic $\mathscr{L}_p$ supports reasoning with statistical and subjective probabilities as two separate entities, and thus has much in common with Halpern's (1990) logic $\mathscr{L}_3$. However, due to the domain distribution semantics of subjective probabilities, $\mathscr{L}_p$ exhibits some distinguishing properties. In this section we will discuss some of these properties. First, however, we turn to purely statistical reasoning, and illustrate by an example the role of the coherence conditions.

Let $\{\mathtt{D}, \mathtt{M}, \ldots\}$ be the vocabulary introduced in section 2.2 for encoding our introductory example. To provide the basis for some inferences in $\mathscr{L}_p$, we first axiomatize some aspects of the intended meaning of the given symbols. Notably, we want $\preceq$ to be an order relation on $\mathtt{M}$, which we can formalize in L_p by a (standard first-order) sentence $\phi_{\preceq}$. Also, according to the intended meaning of $\mathtt{am}$, this function takes values in $\mathtt{M}$:

$$\forall vw(\mathtt{am}(v) = w \rightarrow \mathtt{M}(w)) \equiv: \phi_{\mathtt{am}}.$$

Now consider the statistical probability term

$$[\mathtt{am}(d) \prec \mathtt{am}(d')]_{d,d'}$$

(where $\prec$, naturally, is shorthand for "$\preceq$ and not $=$"), which represents the statistical probability that of two randomly chosen drivers d and d', d has a lower annual mileage than d'. We want to derive that 1/2 is an upper bound for this probability. For this let $\mathfrak{M}$ be a sufficient probabilistic structure for the given vocabulary. Then

$$\begin{aligned} A &:= (\mathfrak{M}, (d, d'))(\mathtt{am}(d) \prec \mathtt{am}(d')) \\ &= \{(a, b) \in M \times M \mid \mathtt{am}(a) \prec \mathtt{am}(b)\} \in \mathfrak{A}_2. \end{aligned} \tag{15}$$

Also, the permutation of A

$$A' := \{(a, b) \in M \times M \mid \mathtt{am}(b) \prec \mathtt{am}(a)\} \tag{16}$$

belongs to $\mathfrak{A}_2$. If $\mathfrak{M}$ is a model of $\phi_{\preceq} \wedge \phi_{\mathtt{am}}$, then A and A' are disjoint, and by homogeneity $P_2(A) = P_2(A')$. It follows that $P_2(A) \leq 1/2$. Hence, we can infer in $\mathscr{L}_p$:

$$\phi_{\preceq} \wedge \phi_{\mathtt{am}} \models [\mathtt{am}(d) \prec \mathtt{am}(d')]_{d,d'} \leq 1/2. \tag{17}$$

Next, we show that from $\phi_{\preceq} \wedge \phi_{\mathtt{am}}$ we can derive

$$\exists d\, [\mathtt{am}(d') \preceq \mathtt{am}(d)]_{d'} \geq 1/2, \tag{18}$$

i.e. there exists a driver whose annual mileage is at least as great as that of 50% of all drivers (an "at least median mileage"-driver). To derive (18) we have to appeal to the Fubini property: let $\mathfrak{M}$ be a model of $\phi_{\preceq} \wedge \phi_{\mathtt{am}}$, and assume that

$$\mathfrak{M} \models \forall d\, [\mathtt{am}(d') \preceq \mathtt{am}(d)]_{d'} < 1/2, \text{i.e.} \tag{19}$$

$$\mathfrak{M} \models \forall d\, [\mathtt{am}(d) \prec \mathtt{am}(d')]_{d'} > 1/2 \tag{20}$$

Now consider again the set A defined by (15). Then, according to (20),

$$A_{1,>1/2} = \{a \in M \mid P_1(\{b \in M \mid a \prec b\}) > 1/2\} = M.$$

By the Fubini property this leads to

$$P_2(A) > 1/2 P_1(M) = 1/2,$$

a contradiction to (17). Hence (20) cannot hold, and (18) follows from $\phi_{\preceq} \wedge \phi_{\mathtt{am}}$.

We now turn to reasoning with subjective probabilities. To simplify notation, we assume in the following that there is only one event symbol e in our vocabulary, i.e. $|\,\mathbf{e}\,| = 1$.

Even though e is interpreted by a probability distribution over the domain, the logic does support the intuition that e, in fact, stands for a unique domain element, because

$$\mathrm{prob}(\exists^{=1} w(\mathrm{e} = w)) = 1 \tag{21}$$

is a tautology in $\mathscr{L}_p$ (here $\exists^{=1}$ is an abbreviation for 'there exists exactly one'). To see that (21) is indeed valid, it only must be realized that the interpretation of the formula $\exists^{=1} w(v = w)$ is always M, and so must be assigned probability 1 by Q_{e}.

Now let $\phi(w)$ be a formula. Then

$$\forall w(\phi(w) \vee \neg\phi(w)) \tag{22}$$

is a tautology. It might now appear as though from (21) and (22) one should be able to infer

$$\phi(\mathrm{e}) \vee \neg\phi(\mathrm{e}), \tag{23}$$

and hence

$$\mathrm{prob}(\phi(\mathrm{e})) = 0 \vee \mathrm{prob}(\phi(\mathrm{e})) = 1. \tag{24}$$

This would mean that reasoning with subjective probabilities reduces to trivial 0-1 valued probability assignments that simply mirror truth value assignments. This is not the case, however, because (23) is an expression that is not allowed by the syntax of $\mathscr{L}_p$, and hence cannot be used for deriving (24). This changes if we introduce a standard constant symbol $\mathtt{e}$ as an alternative name for e via the axiom

$$\mathrm{prob}(\mathrm{e} = \mathtt{e}) = 1. \tag{25}$$

Since $\forall w(w = \mathtt{e} \rightarrow (\phi(w) \leftrightarrow \phi(\mathtt{e})))$ is a tautology, we have

$$\mathrm{prob}(\mathrm{e} = \mathtt{e} \rightarrow (\phi(\mathrm{e}) \leftrightarrow \phi(\mathtt{e})) = 1, \tag{26}$$

and (24) becomes an immediate consequence of (25) and (26).

We thus see that $\mathscr{L}_p$ in this way supports two views on single case probabilities: as long as individual events are only represented by event symbols, the probabilities of their properties can be identified with frequencies obtained by repeated sampling according to Q_{e}, which means that they are only constrained by the conditions of a coherent domain distribution. If the single case nature of e is made explicit by an axiom of the form (25), the logic enforces the view that the probability for a proposition relating to a single case event can only be 0 or 1, according to whether the proposition is true or false. Both these views are shades of frequentist interpretations of single case probabilities: the latter is the strict frequentist view of von Mises (1957), whereas the former is a less dogmatic frequentist perspective in which single case probabilities are admitted as meaningful, but are given an empirical interpretation (Reichenbach 1949, Jaeger 1995*b*).

Limitations on possible subjective probability assignments can be imposed in $\mathscr{L}_p$ also by restricting the sampling distribution Q_{e} in less obvious ways than the axiom (25). Consider the sentence

$$\begin{aligned} &\exists^{=1} v \mathtt{President}(v) \wedge \mathrm{prob}(\mathtt{President}(\mathrm{e})) = 1 \\ &\qquad \wedge \forall v(\mathtt{President}(v) \rightarrow (\mathtt{Republican}(v) \leftrightarrow \neg\mathtt{Democrat}(v))). \end{aligned} \tag{27}$$

The first two conjuncts of this sentence tie the interpretation of e to the one element interpretation of the predicate $\mathtt{President}$ in very much the same way as (25) tied it to the one element interpretation of $\mathtt{e}$. As before, we thus obtain that properties of e can only have 0-1 probabilities, and hence (27) is inconsistent with

$$\mathrm{prob}(\mathtt{Republican}(\mathrm{e})) = 1/2 \wedge \mathrm{prob}(\mathtt{Democrat}(\mathrm{e})) = 1/2. \tag{28}$$

This may seem counterintuitive at first sight, as (27) and (28) seem to express a meaningful subjective probability assessment. On the other hand, however, it also seems natural to demand that for any formula $\phi(x)$ the implication

$$\mathrm{prob}(\phi(\mathrm{e})) > 0 \models \exists v \phi(v) \tag{29}$$

should be valid, since we should not be able to assign a nonzero probability to e having the impossible property ϕ. If, now, (27) and (28) were jointly consistent, then (29) would be violated in some model with either $\phi(v) = \mathtt{President}(v) \wedge \mathtt{Democrat}(v)$, or $\phi(v) = \mathtt{President}(v) \wedge \mathtt{Republican}(v)$. Thus, the minimal consistency requirement between domain knowledge and subjective probability assessment as expressed by (29) already forces the joint inconsistency of (27) and (28).

A somewhat more careful modeling resolves the apparent conflict: by introducing a time parameter into our representation, we can make the more accurate statement that there only exists a single president at any given point in time, and that e refers to the next president:

$$\forall t \texttt{Time}(t) \rightarrow \exists^{=1} v \texttt{President}(v,t) \wedge \mathrm{prob}(\texttt{President}(\mathrm{e}, \mathrm{next})) = 1. \tag{30}$$

Here 'next' must be another event, not a constant symbol. Now (28) is consistent with our premises since $Q_{\mathrm{e,next}}$ can be any distribution that samples presidents at different points in time.

2.5 Sufficient Structures

So far, the only type of sufficient probabilistic structures we have encountered are the real-discrete structures of example 2.7. For many interesting theories one can find models that belong to this class. For instance, all our example sentences (1),(3), etc. have real discrete models. This is not always the case, though. Consider the sentence

$$\phi^{\mathrm{cont}} :\equiv \forall v[v = w]_w = 0,$$

which explicitly states that no single element carries a positive probability mass. Clearly ϕ^{cont} does not have a real discrete model. Probabilistic structures that do satisfy ϕ^{cont} we call *continuous structures*. Do sufficient continuous structures exist? The answer is yes. An explicit construction of sufficient continuous structures for the special case that S only contains unary relation symbols is given in (Jaeger 1995*a*). For more expressive vocabularies it becomes extremely difficult to verify sufficiency in an explicit construction. In particular, as the following theorem shows, we cannot follow the example of real-discrete structures, and try to obtain sufficiency simply by making every set measurable.

Theorem 2.8 There does not exist a sufficient continuous probability structure $\mathfrak{M}$ with $\mathfrak{A}_n = 2^{M^n}$ for all n.

Proof: We show the stronger result that we cannot even construct the first two elements $(2^M, P_1), (2^{M^2}, P_2)$ of a sequence $(2^{M^n}, P_n)_{n \in \mathbb{N}}$ such that the coherency conditions hold for these two measure algebras.

For this let M be a set, P_1 a continuous probability measure on 2^M, P_2 a permutation invariant probability measure on 2^{M^2} such that P_1 and P_2 satisfy the product property. We show that there exists an $A \subseteq M^2$ with $P_1(\sigma^1_a(A)) = 0$ for all $a \in M$, and $P_2(A) > 0$, thus providing a counterexample to the Fubini property.

Let λ be the cardinality of M. Let Γ be the set of ordinals $\kappa \leq \lambda$ that have the following property: there exists a sequence of pairwise disjoint subsets $\{E_\nu \subset M \mid \nu \text{ ordinal}, \ \nu < \kappa\}$ with

$$\forall \nu < \kappa: \ P_1(E_\nu) = 0 \ \text{ and } \ P_1(\cup_{\nu<\kappa} E_\nu) > 0. \tag{31}$$

Γ is nonempty, because $\lambda \in \Gamma$.

Let ρ be the minimal element in Γ; let $\{E_\nu \mid \nu < \rho\}$ be a sequence for ρ with (31). For each ordinal $\nu < \rho$ let

$$\tilde{E}_\nu := \cup_{\theta < \nu} E_\theta.$$

By the minimality of ρ in Γ, we have $P_1(\tilde{E}_\nu) = 0$ for all $\nu < \rho$. Now define

$$\begin{aligned} A_0 &:= \cup_{\nu<\rho}(E_\nu \times \tilde{E}_\nu), \\ A_1 &:= \cup_{\nu<\rho}(E_\nu \times E_\nu), \\ B &:= \cup_{\nu<\rho} E_\nu. \end{aligned}$$

Let $a \in M$ be arbitrary. If $a \notin B$, then $\sigma_a^1(A_0) = \sigma_a^1(A_1) = \emptyset$. For $a \in B$ there exists exactly one $\nu < \rho$ with $a \in E_\nu$, so that $\sigma_a^1(A_0) = \tilde{E}_\nu$ and $\sigma_a^1(A_1) = E_\nu$. Thus, for all $a \in M$, $P_1(\sigma_a^1(A_0)) = P_1(\sigma_a^1(A_1)) = 0$.

Now consider any $(a, b) \in B \times B$ where $a \in E_\nu$, $b \in E_{\nu'}$. If $\nu > \nu'$ then $(a, b) \in A_0$. For $\nu = \nu'$ we have $(a, b) \in A_1$, and if $\nu < \nu'$, then (a, b) belongs to the permutation $\pi A_0 := \cup_{\nu<\rho}(\tilde{E}_\nu \times E_\nu)$ of A_0. Thus,

$$B \times B = A_0 \cup \pi A_0 \cup A_1.$$

Since $r := P_1(B) > 0$, and therefore $P_2(B \times B) = r^2 > 0$, by the permutation invariance of P_2, it follows that $P_2(A_0) > 0$, or $P_2(A_1) > 0$. Hence, at least one of A_0 and A_1 violates the Fubini property. □

2.6 Reduction to First-Order Logic

The previous section has highlighted the difficulties in the model theory of $\mathscr{L}_p$. In this section we provide results that, on the other hand, provide powerful tools for the analysis of $\mathscr{L}_p$. These tools are obtained by showing that $\mathscr{L}_p$ can be reduced to standard first-order logic. This reduction is based on the observation that a statistical probability term $[\phi(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{x})]_{\boldsymbol{w}}$ maps tuples $(\boldsymbol{a}, \boldsymbol{r}) \in M^{|\boldsymbol{v}|} \times \mathbb{F}^{|\boldsymbol{x}|}$ to elements $s \in \mathbb{F}$, and thus behaves essentially like a standard function term $\mathtt{f}(\boldsymbol{v}, \boldsymbol{x})$ over a domain $M \cup \mathbb{F}$. A similar observation applies to subjective probability terms. To reduce $\mathscr{L}_p$ to first-order logic, one can define a translation from L_p into the language $L_I(S^*)$ of first-order logic over an expanded (infinite) vocabulary $S^* \supset S$. In this translation, probability terms are inductively replaced by standard function terms using new function symbols. This syntactic translation is complemented by a transformation between sufficient probabilistic structures and standard first-order structures. Finally, the class of standard first-order structures that correspond to sufficient probabilistic structures under such a transformation can be axiomatized by a first-order theory AX. We then obtain the following result.

Theorem 2.9 Let S be a vocabulary. There exist

- a vocabulary $S^* \supset S$,
- a recursively enumerable set of axioms $\mathrm{AX} \subset L_I(S^*)$,
- computable mappings

$$\begin{aligned} t &: L_p(S) \rightarrow L_I(S^*) \\ t^{-1} &: t(L_p(S)) \rightarrow L_p(S), \end{aligned}$$

  such that $t^{-1}(t(\phi)) = \phi$,

- transformations

$$\begin{array}{ll} T : \mathfrak{M} \mapsto \mathfrak{M}^* & (\mathfrak{M} \text{ a sufficient probabilistic } S\text{-structure}, \\ & \ \mathfrak{M}^* \text{ a } S^*\text{-structure with } \mathfrak{M}^* \models \mathrm{AX}) \\ T^{-1} : \mathfrak{N} \rightarrow \mathfrak{N}^{-1} & (\mathfrak{N} \text{ a } S^*\text{-structure with } \mathfrak{N} \models \mathrm{AX}, \\ & \ \mathfrak{N}^{-1} \text{ a sufficient probabilistic } S\text{-structure})\ , \end{array}$$

  such that $T^{-1}(T(\mathfrak{M})) = \mathfrak{M}$,

so that for all $\phi \in L_p(S)$, all sufficient probabilistic S-structures $\mathfrak{M}$, and all S^*-structures $\mathfrak{N} \models \mathrm{AX}$:

$$\mathfrak{M} \models \phi \ \text{ iff } \ T(\mathfrak{M}) \models t(\phi) \quad \text{and} \quad \mathfrak{N} \models t(\phi) \ \text{ iff } \ T^{-1}(\mathfrak{N}) \models \phi. \tag{32}$$

For the detailed proof of this theorem the reader is referred to (Jaeger 1995*a*). We obtain several useful corollaries. The first one reduces semantic implication in $\mathscr{L}_p$ to first-order entailment.

Corollary 2.10 For all $\Phi \cup \{\phi\} \subseteq L_p(S)$:

$$\Phi \models \phi \ \text{ iff } \ t(\Phi) \cup \mathrm{AX} \models t(\phi).$$

Using this corollary, one can easily transfer compactness of first-order logic to $\mathscr{L}_p$:

Corollary 2.11 $\mathscr{L}_p$ is compact.

As an application of compactness consider the L_p-theory

$$\Phi := \ \{\delta_n \mid n \in \mathbb{N}\} \cup \ \exists x > 0 \forall v [v = w]_w = x,$$

where δ_n is a standard first-order sentence that says that the domain contains at least n elements. A model of Φ thus is an infinite structure in which every singleton has the same positive probability mass. Since every finite subset of Φ is satisfiable (by a finite domain real-discrete structure), we know by corollary 2.11 that Φ is satisfiable. However, Φ is clearly not satisfiable by a structure with real-valued probabilities: the probability of the singletons in a model of Φ must be some infinitesimal. Thus, Φ also provides an example of what we lose in terms of semantic strength by allowing probabilities to be lrc-field-valued, not necessarily real-valued, and shows that corollary 2.11 cannot hold when we limit ourselves to real-valued probability structures.

Finally, we obtain as a corollary to theorem 2.9 a completeness result:

Theorem 2.12 There exists a sound and complete proof system for $\mathscr{L}_p$.

Again, this corollary is in marked contrast to what one obtains when probabilities are required to be real-valued, in which case no complete proof system can exist (Abadi & J.Y.Halpern 1994).

3 The Logic of Inductive Inference

3.1 Inductive Reasoning by Cross-Entropy Minimization

The statistical knowledge expressed in our example sentences (1),(4) and (6) can be expressed by the L_p-sentences

$$\begin{aligned}
\phi_1 &:\equiv [\mathtt{IIA}(d) \mid \mathtt{D}(d) \wedge 10 \preceq \mathtt{am}(d) \preceq 20]_d = 0.027 \qquad (33)\\
\phi_4 &:\equiv [\mathtt{IIA}(d) \mid \mathtt{D}(d) \wedge 15 \preceq \mathtt{am}(d) \preceq 25]_d = 0.031 \qquad (34)\\
\phi_6 &:\equiv [\mathtt{IIA}(d) \mid \mathtt{D}(d) \wedge 15 \preceq \mathtt{am}(d) \preceq 20]_d \in [0.027, 0.031]. \qquad (35)
\end{aligned}$$

The belief about Jones expressed in (2) can be expressed by

$$\phi_2 :\equiv \mathrm{prob}(\mathtt{D}(\mathit{jones}) \wedge 10 \preceq \mathtt{am}(\mathit{jones}) \preceq 20) = 1. \tag{36}$$

As discussed in the introduction, it seems reasonable to infer from $\phi_1 \wedge \phi_2$

$$\phi_3 :\equiv \mathrm{prob}(\mathtt{IIA}(\mathit{jones})) = 0.027. \tag{37}$$

However, this inference is not valid in $\mathscr{L}_p$, i.e.

$$\phi_1 \wedge \phi_2 \not\models \phi_3.$$

This is because in a probabilistic structure the statistical and subjective probability terms are interpreted by the measures P_1 and Q_{jones}, respectively, and the constraint ϕ_1 on admissible statistical measures does not constrain the possible choices for Q_{jones}. Moreover, it would clearly not be desirable to have that $\phi_1 \wedge \phi_2$ strictly implies ϕ_3, because then $\phi_1 \wedge \phi_2$ would be inconsistent with $\mathrm{prob}(\neg\mathtt{IIA}(\mathit{jones})) = 1$, i.e. the knowledge that Jones will, in fact, not be involved in an accident. Hence, if we wish to infer ϕ_3 from $\phi_1 \wedge \phi_2$, this can only have the character of a *nonmonotonic*, or *defeasible*, inference, which may become invalid when additional information becomes available. Our aim, then, will be to augment the logic $\mathscr{L}_p$ with an additional nonmonotonic entailment relation $\mathrel{|\!\approx}$ for which

$$\phi_1 \wedge \phi_2 \mathrel{|\!\approx} \phi_3, \text{ but } \phi_1 \wedge \phi_2 \wedge \mathrm{prob}(\neg\mathtt{IIA}(\mathit{jones})) = 1 \not\mathrel{|\!\approx} \phi_3.$$

As a second example for the intended inference relation $\mathrel{|\!\approx}$ consider the formula

$$\phi_{2,5} :\equiv \mathrm{prob}(\mathtt{D}(\mathit{jones}) \wedge 15 \preceq \mathtt{am}(\mathit{jones}) \preceq 20) = 1. \tag{38}$$

As argued in the introduction, our inductive inference relation then should satisfy

$$\phi_6 \wedge \phi_{2,5} \mathrel{|\!\approx} \mathrm{prob}(\mathtt{IIA}(\mathit{jones})) \in [0.027, 0.031].$$

Adding that these should be the sharpest bounds that $\mathrel{|\!\approx}$ allows us to derive for $\mathrm{prob}(\mathtt{IIA}(\mathit{jones}))$, this example illustrates an important aspect of the intended relation $\mathrel{|\!\approx}$: it will not be used to make any default assumptions about the statistical distribution in the sense that, for example, we could derive

$$\phi_6 \mathrel{|\!\approx} [\mathtt{IIA}(d) \mid \mathtt{D}(d) \wedge 15 \preceq \mathtt{am}(d) \preceq 20]_d = 0.029$$

(i.e. assuming without further information that the correct statistical probability is given by the center point of the admissible interval, or else, maybe, by 0.031 as the value closest to 0.5). Only inferring the bounds $[0.027, 0.031]$ for $\text{prob}(\texttt{IIA}(\textit{jones}))$ means that we take every admissible statistical distribution into consideration, and apply the inductive inference relation $\mid\!\approx$ to the subjective distribution alone with respect to each of the statistical possibilities.

As an example where the given information on Jones is not deterministic consider the sentence

$$\begin{aligned}\phi_{39} :\equiv\ &\text{prob}(\texttt{D}(\textit{jones}) \wedge 10 \preceq \texttt{am}(\textit{jones}) \preceq 15) = 0.4\\ &\wedge\ \text{prob}(\texttt{D}(\textit{jones}) \wedge 15 \preceq \texttt{am}(\textit{jones}) \preceq 20) = 0.6. \qquad (39)\end{aligned}$$

Here Jeffrey's rule is applicable, because the two constraints in (39) are on disjoint subsets. Jeffrey's rule, now, leads to the inductive inference

$$\begin{aligned}\phi_{39} \mid\!\approx \text{prob}(\texttt{IIA}(\textit{jones})) =&0.4[\texttt{IIA}(d) \mid \texttt{D}(d) \wedge 10 \preceq \texttt{am}(d) \preceq 15]_d +\\ &0.6[\texttt{IIA}(d) \mid \texttt{D}(d) \wedge 15 \preceq \texttt{am}(d) \preceq 20]_d.\end{aligned} \qquad (40)$$

As the statistical information $\phi_1 \wedge \phi_6$ implies the bounds $[0, 0.027]$ and $[0.027, 0.031]$ for the two conditional probabilities on the right hand side of (40), we obtain

$$\begin{aligned}\phi_1 \wedge \phi_6 \wedge \phi_{39} \mid\!\approx \text{prob}(\texttt{IIA}(\textit{jones})) &\in [0.6 \cdot 0.027, 0.4 \cdot 0.027 + 0.6 \cdot 0.031]\\ &= [0.0162, 0.0294]. \qquad (41)\end{aligned}$$

While the step from direct inference to Jeffrey's rule is very easy, the step to the general case where subjective probability constraints can be on arbitrary, non-disjoint, sets is rather non-trivial. The guiding principle both in direct inference and Jeffrey's rule can be seen as the attempt to make the subjective probability distribution as similar as possible to the statistical distribution. To follow this principle in general requires to be able to measure the similarity, or distance, between probability distributions. A very prominent distance measure for probability distributions is *cross-entropy*: if $P = (p_1, \dots, p_n)$ and $Q = (q_1, \dots, q_n)$ are two probability measures on an n-element probability space, and $p_i = 0$ implies $q_i = 0$ for $i = 1, \dots, n$ (i.e. Q is *absolutely continuous* with respect to Q, written $Q \ll P$), then the cross-entropy of Q with respect to P is defined by

$$CE(Q, P) := \sum_{\substack{i=1\\ p_i>0}}^{n} q_i Log \frac{q_i}{p_i}. \qquad (42)$$

Given a measure $P \in \Delta\mathfrak{A}$ with $\mathfrak{A}$ a finite algebra, and a subset $J \subseteq \Delta\mathfrak{A}$, we can define the *CE-projection* of P onto J:

$$\Pi_J(P) := \{Q \in J \mid Q \ll P,\ \forall Q' \in J :\ CE(Q', P) \geq CE(Q, P)\}. \qquad (43)$$

The set $\Pi_J(P)$ can be empty (either because J does not contain any Q with $Q \ll P$, or because the infimum of $\{CE(Q', P) \mid Q' \in J\}$ is not attained by any $Q \in J$), can be a singleton, or contain more than one element.

To use CE in modeling inductive probabilistic reasoning, we identify the distributions P and Q in (42) with the statistical and subjective probability distributions, respectively. We can then formalize the process of inductive probabilistic reasoning as follows: if K is the set of statistical measures consistent with our knowledge, J is the set of subjective measures consistent with our already formed, partial beliefs, then we will sharpen our partial beliefs by going from J to

$$\Pi_J(K) := \cup\{\Pi_J(P) \mid P \in K\} \ \subseteq J,$$

i.e. by discarding all subjective distributions that are not as close as possible to at least one feasible statistical distribution.

Is this an adequate formalization of inductive probabilistic reasoning? Clearly, this question, being non-mathematical in nature, does not admit of an affirmative answer in the form of a strict correctness proof. However, it is arguable that, short of such a proof, the justification for using cross-entropy minimization is as strong as it possibly can be.

A first justification consists in the observation that cross-entropy minimization does indeed generalize Jeffrey's rule: if J is defined by prescribing values for the elements of a partition, then $\Pi_J(P)$ is obtained by applying Jeffrey's rule to P and these values. This property, however, is not unique to cross-entropy minimization (Diaconis & Zabell 1982). Justifications that identify cross-entropy minimization as the unique method satisfying certain desirable properties can be brought forward along two distinct lines: the first type of argument consists of formal conditions on the input/output relation defined by a method, and a proof that cross-entropy minimization is the only rule that will satisfy these conditions. This approach underlies the well-known works both by Shore and Johnson (1980, 1983) and of Paris and Vencovská (1990, 1992). A typical condition that will be postulated in derivations of this type can be phrased in terms of inductive inference in L_p as follows: if the input consists of separate constraints on two event variables, e.g.

$$\mathrm{prob}(10 \preceq \mathtt{am}(\mathit{jones}) \preceq 15) \leq 0.7 \ \wedge \ \mathrm{prob}(\mathtt{IIA}(\mathit{mitchell})) \leq 0.1, \tag{44}$$

then the output, i.e. the selected joint subjective distribution for Jones and Mitchell, should make the two variables independent, and therefore satisfy e.g.

$$\begin{aligned} &\mathrm{prob}(\mathtt{IIA}(\mathit{jones}) \wedge 10 \preceq \mathtt{am}(\mathit{mitchell})) \\ &\qquad = \mathrm{prob}(\mathtt{IIA}(\mathit{jones})) \cdot \mathrm{prob}(10 \preceq \mathtt{am}(\mathit{mitchell})). \end{aligned} \tag{45}$$

Abstracting from such particular examples, this independence principle becomes a general property of the inductive entailment operator $\approx\!\!\!\!|$, which can be formally stated as in theorem 3.8 below (and which corresponds to the *system independence* property of (Shore & Johnson 1980), respectively the principle of independence of (Paris 1994)). A second condition, or desideratum, for an inductive inference rule is the conditional reasoning property, expressed in theorem 3.9 below (which is closely related to the *subset independence* property of (Shore & Johnson 1980)). Variants of these two properties form the core of axiomatic derivations of *CE*-minimization as the formal rule for inductive probabilistic inference.

A second type of justification for the minimum *CE*-principle has been developed in (Jaeger 1995*b*, Jaeger 1995*a*). This justification follows the tradition of frequentist interpretations for single case probabilities as predicted frequencies in a sequence of trials (Reichenbach 1949, §72),(Carnap 1950, p. 189ff).

Since single case probabilities often cannot be associated with observable frequencies in actual, repeated, physical experiments, such trials may only take an imaginary form, i.e. be carried out as a thought experiment (Jaeger 1995*b*). For example, to assess the probability that the driver of the car, the wreckage of which we have just seen at the roadside, has survived the crash, we may mentally reenact the accident several times, and take a mental count of how often the driver comes away alive. We now make two assumptions about how the thought experiment is performed. The first assumption is that the sampling in the thought experiment is according to our statistical knowledge of the domain. If, for example, we happen to know exact statistics on the average speed of vehicles on this road, the prevalence of seat-belt use, the frequency of drunk driving, etc., then our mental sampling will be in accordance with these known statistics. The second assumption is that already existing constraints on the subjective probability being assessed are used to condition the statistical distribution over possible samples on frequencies consistent with these constraints. If, for example, we happen to believe that with probability at least 0.7 the driver in the accident was drunk (this being well above the statistical probability of drunk driving), then we condition the distribution over possible samples of repeated accidents on the event of containing at least 70% incidences of drunk driving. More loosely speaking, we perform the mental sampling according to the underlying statistical distribution, but bias the result so as to contain at least 70% drunk drivers.

This semi-formal thought experiment model can be translated into a precise statistical model, and it can then be proven that according to this model the predicted frequencies must be exactly those that are obtained by *CE*-minimization (Jaeger 1995*b*).

As an example for a result obtained by *CE*-minimization in a situation where Jeffrey's rule no longer applies, consider the sentence

$$\begin{aligned}\phi_{46} :\equiv \;& \text{prob}(10 \preceq \texttt{am}(jones) \preceq 20) = 0.5 \\ & \wedge\text{prob}(15 \preceq \texttt{am}(jones) \preceq 25) = 0.7,\end{aligned} \tag{46}$$

This sentence imposes probability constraints on the two non-disjoint sets defined by $10 \preceq \texttt{am}(v) \preceq 20$ and $15 \preceq \texttt{am}(v) \preceq 25$. As usual, we want to derive a probability estimate for $\texttt{IIA}(jones)$. It is another distinctive feature of *CE*-minimization that this estimate can be derived in two steps as follows: in the first step probability estimates for Jones belonging to the elements of the partition generated by the sets appearing in (46) are computed (by *CE*-minimization). In the second step the probability assignments found for the partition are extended to other sets by Jeffrey's rule, which now is applicable. For example ϕ_{46} the relevant partition consists of four different sets of possible annual mileages, for which we might have the following statistical information:

$$[10 \preceq \mathtt{am}(d) \prec 15]_d = 0.4 \tag{47}$$

$$[15 \preceq \mathtt{am}(d) \preceq 20]_d = 0.3 \tag{48}$$

$$[20 \prec \mathtt{am}(d) \preceq 25]_d = 0.1 \tag{49}$$

$$[\mathtt{am}(d) \prec 10 \vee 25 \prec \mathtt{am}(d)]_d = 0.2 \tag{50}$$

To obtain the probability estimates for Jones's membership in the elements of the partition, we have to compute the distribution $Q = (q_1, q_2, q_3, q_4)$ that minimizes $CE(\cdot, P)$ with respect to $P = (0.4, 0.3, 0.1, 0.2)$ under the constraints $q_1 + q_2 = 0.5$ and $q_2 + q_3 = 0.7$. This computation is a non-linear optimization problem, and yields the (approximate) solution

$$Q = (0.128\ldots, 0.37\ldots, 0.329\ldots, 0.171\ldots), \tag{51}$$

meaning that in the first step we have made, for example, the inductive inference

$$\text{prob}(10 \preceq \mathtt{am}(\mathit{jones}) \preceq 15) \in (0.128, 0.129). \tag{52}$$

Given the probabilities for the four disjoint reference classes we can now apply Jeffrey's rule, and obtain bounds for prob($\mathtt{IIA}(\mathit{jones})$) in the same way as (41) was derived from (39) and the relevant statistical information.

3.2 Preferred Models

Having identified cross-entropy minimization as the formal rule we want to employ for inductive reasoning, we want to use it as the basis for inductive entailment $\mid\!\approx$ in $\mathscr{L}_p$.

Our plan is to implement *CE*-minimization by developing a *preferred model semantics* (Shoham 1987) for L_p: for a given L_p-sentence ϕ we will single out from the set of all models of ϕ a subset of preferred models. A model $\mathfrak{M} = (M, \ldots, (\mathfrak{A}_n, P_n)_n, Q_{\mathbf{e}})$ is going to be a preferred model if the subjective probability measure $Q_{\mathbf{e}}$ minimizes cross-entropy with respect to the measure $P_{|\mathbf{e}|}$ that describes the statistical distribution of a random sample of $|\,\mathbf{e}\,|$ domain elements. An inductive entailment relation $\phi \mid\!\approx \psi$ then holds if ψ is true in all preferred models of ϕ.

Several difficulties arise when we put this plan into practice, because, we have defined cross-entropy by (42) only for real-valued measures on finite algebras. As we are now dealing with lrc-field valued measures on infinite algebras, the concepts of cross-entropy and *CE*-minimization have to be generalized. Furthermore, we have to ascertain that this generalization retains those essential properties of cross-entropy in $\mathbb{R}$ on which the justification of the minimum *CE*-principle is based. For instance, we will have to check that the generalized minimum *CE*-principle still has the independence property, so that the inductive inference of (45) from (44) remains valid with our lrc-field based semantics.

We tackle this complex of questions in two stages: first we define cross-entropy for lrc-field valued measures on finite spaces, and prove that here generalized cross-entropy exhibits the same essential properties as cross-entropy on the reals. In a second step we show that for our purpose it is already sufficient to define cross-entropy on finite algebras, because a suitable notion of *CE*-minimization for measures on the infinite algebra $\mathfrak{A}_{|\mathbf{e}|}$ can be obtained by "lifting" cross-entropy minimal measures from finite subalgebras of $\mathfrak{A}_{|\mathbf{e}|}$ to $\mathfrak{A}_{|\mathbf{e}|}$.

To begin, we have to define cross-entropy and *CE*-projections for lrc-field valued measures on finite algebras. This, however, is immediate, and is done by (42) and (43) just as for real-valued measures simply by interpreting the function *Log* now as an arbitrary logarithmic function in an lrc-field.

This leads us to the question of what properties of cross-entropy in the reals carry over to the generalized *CE* function. We give a fairly comprehensive answer to this question in appendix A: first we show that *CE*-projections in lrc-fields retain the key structural properties of *CE*-projections in the reals, namely those properties on which Shore and Johnson (1980) base their derivation of the minimum *CE*-principle. From these results it follows, for example, that the inductive inference from (44) to (45) also is warranted on the basis of lrc-field valued probabilities. Second, it is shown in appendix A that generalized *CE*-minimization also behaves numerically essentially as *CE*-minimization in the reals. This means, for example, that the numerical result (52) also is obtained with lrc-field valued probabilities. In summary, the results developed in appendix A constitute a collection of far-reaching completeness results that show that for finite algebras we retain for *CE*-minimization in lrc-fields most of the salient features of *CE*-minimization for real-valued measures. In some of the proofs of theorems in the present section references are made to results of appendix A. It should be noted that all these references are to facts that are long established for real-valued probabilities, and therefore are inessential as long as one follows the main development thinking of real-valued probabilities alone.

It remains to find a suitable notion of *CE*-minimization for measures defined on $\mathfrak{A}_{|\mathbf{e}|}$ by a reduction to *CE*-minimization on finite algebras. Although the following construction contains some technicalities, the underlying idea is extremely simple, and consists essentially of the same two-step procedure used in the example (46)-(52) of the preceding section. To be able to carry out the first step of that procedure, it is necessary that the given constraints on the subjective distribution only refer to finitely many sets, which will generate a finite partition on which we know how to conduct *CE*-minimization. In the following we give a precise semantic definition for what it means that constraints only refer to finitely many sets. Later (lemma 3.6) we will see that constraints expressible in $\mathscr{L}_p$ are guaranteed to have this semantic property.

Definition 3.1 Let $\mathfrak{A}$ be an algebra over M. Let $J \subseteq \Delta_{\mathbb{F}}\mathfrak{A}$, and $\mathfrak{A}'$ a finite subalgebra of $\mathfrak{A}$. Let $J \upharpoonright \mathfrak{A}' := \{P \upharpoonright \mathfrak{A}' \mid P \in J\}$. We say that J *is defined by constraints on* $\mathfrak{A}'$, iff

$$\forall P \in \Delta_{\mathbb{F}}\mathfrak{A} : \quad P \in J \;\text{ iff }\; P \upharpoonright \mathfrak{A}' \in J \upharpoonright \mathfrak{A}'.$$

Given a set $J \subseteq \Delta_{\mathbb{F}}\mathfrak{A}$ defined by constraints on some finite $\mathfrak{A}' \subseteq \mathfrak{A}$, we can apply the two step process of first computing $\Pi_{J \upharpoonright \mathfrak{A}'}(P \upharpoonright \mathfrak{A}')$, and then extend the result to $\mathfrak{A}$ by Jeffrey's rule as formally described in the following definition.

Definition 3.2 Let $\mathfrak{A}$ be an algebra, $P \in \Delta_{\mathbb{F}}\mathfrak{A}$. Let $\mathfrak{A}' \subseteq \mathfrak{A}$ a finite subalgebra with atoms $\{A_1, \ldots, A_L\}$, and $Q \in \Delta_{\mathbb{F}}\mathfrak{A}'$ such that $Q \ll P \upharpoonright \mathfrak{A}'$. Let P^h be the conditional distribution of P on A_h ($h = 1, \ldots, L;\ P(A_h) > 0$). The extension Q^* of Q to $\mathfrak{A}$ defined by

$$Q^* := \sum_{\substack{h=1 \\ P(A_h)>0}}^{L} Q(A_h) P^h$$

is called the *Jeffrey-extension* of Q to $\mathfrak{A}$ by P, denoted by $\mathscr{J}(Q, P, \mathfrak{A})$.

The following lemma says that if J is defined by constraints on $\mathfrak{A}'$, then Jeffrey extensions realize cross-entropy minimization on all finite algebras that refine $\mathfrak{A}'$.

Lemma 3.3 Let $\mathfrak{A}$ be an algebra, $P \in \Delta_{\mathbb{F}}\mathfrak{A}$. Let $J \subseteq \Delta_{\mathbb{F}}\mathfrak{A}$ be defined by constraints on a finite subalgebra $\mathfrak{A}' \subseteq \mathfrak{A}$. Then for all finite $\mathfrak{A}'' \supseteq \mathfrak{A}'$:

$$\Pi_{J \upharpoonright \mathfrak{A}''}(P \upharpoonright \mathfrak{A}'') = \{Q \upharpoonright \mathfrak{A}'' \mid Q = \mathscr{J}(Q', P, \mathfrak{A}),\ Q' \in \Pi_{J \upharpoonright \mathfrak{A}'}(P \upharpoonright \mathfrak{A}')\}. \tag{53}$$

Conversely, for $Q \in \Delta_{\mathbb{F}}\mathfrak{A}$, if

$$Q \upharpoonright \mathfrak{A}'' \in \Pi_{J \upharpoonright \mathfrak{A}''}(P \upharpoonright \mathfrak{A}'') \tag{54}$$

for all finite $\mathfrak{A}'' \supseteq \mathfrak{A}'$, then $Q = \mathscr{J}(Q \upharpoonright \mathfrak{A}', P, \mathfrak{A})$.

Proof: Let $\{A_1, \ldots, A_p\}$ be the set of atoms of $\mathfrak{A}'$. Let $Q'' \in \Delta_{\mathbb{F}}\mathfrak{A}''$, $Q'' \ll P \upharpoonright \mathfrak{A}''$. By lemma A.2 then

$$CE(Q'', P \upharpoonright \mathfrak{A}'') \geq CE(Q'' \upharpoonright \mathfrak{A}', P \upharpoonright \mathfrak{A}')$$

with equality iff

$$(Q'')^h = (P \upharpoonright \mathfrak{A}'')^h \qquad (h = 1, \ldots, p) \tag{55}$$

where $(\cdot)^h$ is the conditional distribution on A_h. Equivalent to (55) is

$$Q'' = \mathscr{J}(Q'' \upharpoonright \mathfrak{A}', P \upharpoonright \mathfrak{A}'', \mathfrak{A}'').$$

Since J is defined by constraints on $\mathfrak{A}'$, we have for all $Q' \in J \upharpoonright \mathfrak{A}'$ that

$$\mathscr{J}(Q', P \upharpoonright \mathfrak{A}'', \mathfrak{A}'') \in J \upharpoonright \mathfrak{A}'',$$

and therefore

$$\Pi_{J \upharpoonright \mathfrak{A}''}(P \upharpoonright \mathfrak{A}'') = \{\mathscr{J}(Q', P \upharpoonright \mathfrak{A}'', \mathfrak{A}'') \mid Q' \in \Pi_{J \upharpoonright \mathfrak{A}'}(P \upharpoonright \mathfrak{A}')\}. \tag{56}$$

With

$$\mathscr{J}(Q', P \upharpoonright \mathfrak{A}'', \mathfrak{A}'') = \mathscr{J}(Q', P, \mathfrak{A}) \upharpoonright \mathfrak{A}''$$

this proves (53).

Conversely, assume that (54) holds for Q and all finite $\mathfrak{A}''$. Then, in particular, $Q \upharpoonright \mathfrak{A}' \in \Pi_{J \upharpoonright \mathfrak{A}'}(P \upharpoonright \mathfrak{A}')$, and, again by lemma A.2,

$$Q \upharpoonright \mathfrak{A}'' = \mathscr{J}(Q \upharpoonright \mathfrak{A}', P \upharpoonright \mathfrak{A}'', \mathfrak{A}'')$$

for all finite $\mathfrak{A}'' \supseteq \mathfrak{A}'$. Thus, also $Q = \mathscr{J}(Q \upharpoonright \mathfrak{A}', P, \mathfrak{A})$. □

Lemma 3.3 suggests to define for $J \subseteq \Delta\mathfrak{A}$ that is defined by constraints on the finite subalgebra $\mathfrak{A}' \subseteq \mathfrak{A}$:

$$\Pi_J(P) := \{\mathscr{J}(Q', P, \mathfrak{A}) \mid Q' \in \Pi_{J \upharpoonright \mathfrak{A}'}(P \upharpoonright \mathfrak{A}')\} \tag{57}$$

However, there is still a slight difficulty to overcome: the algebra $\mathfrak{A}'$ is not uniquely determined, and (57) would be unsatisfactory if it depended on the particular choice of $\mathfrak{A}'$. We therefore show, next, that this is not the case, which is basically due to the fact that there is a unique smallest algebra $\mathfrak{A}'$ by constraints on which J is defined.

Lemma 3.4 Let $\mathfrak{A}$ be an algebra, $\mathfrak{A}'$ and $\mathfrak{A}''$ finite subalgebras of $\mathfrak{A}$. Assume that $J \subseteq \Delta\mathfrak{A}$ is defined by constraints on $\mathfrak{A}'$, and also by constraints on $\mathfrak{A}''$. Then J also is defined by constraints on

$$\mathfrak{A}^{\cap} := \mathfrak{A}' \cap \mathfrak{A}''.$$

Proof: Let $\mathfrak{A}^{\cup}$ be the subalgebra of $\mathfrak{A}$ generated by $\mathfrak{A}'$ and $\mathfrak{A}''$. Then J also is defined by constraints on $\mathfrak{A}^{\cup}$, and it suffices to show that for all $Q \in \Delta\mathfrak{A}$

$$Q \upharpoonright \mathfrak{A}^{\cup} \in J \upharpoonright \mathfrak{A}^{\cup} \quad \Leftrightarrow \quad Q \upharpoonright \mathfrak{A}^{\cap} \in J \upharpoonright \mathfrak{A}^{\cap}. \tag{58}$$

To obtain a more economical notation, we may therefore work within a completely finitary context, and assume that $\mathfrak{A} = \mathfrak{A}^{\cup}$ and $J \subseteq \Delta_{\mathbb{F}}\mathfrak{A}^{\cup}$.

With $\{A'_i \mid i = 1, \ldots, p\}$ the atoms of $\mathfrak{A}'$, and $\{A''_j \mid j = 1, \ldots, q\}$ the atoms of $\mathfrak{A}''$, atoms of $\mathfrak{A}^{\cup}$ are the nonempty intersections

$$B_{ij} := A'_i \cap A''_j \;\; (i = 1, \ldots, p; \; j = 1, \ldots, q).$$

Elements of $\mathfrak{A}^\cap$ are just the unions of atoms of $\mathfrak{A}'$ that simultaneously can be represented as a union of atoms of $\mathfrak{A}''$, i.e.

$$A = \bigcup_{i \in I} A'_i \in \mathfrak{A}'$$

with $I \subseteq \{1, \ldots, p\}$ belongs to $\mathfrak{A}^\cap$ iff there exists $K \subseteq \{1, \ldots, q\}$, such that also

$$A = \bigcup_{k \in K} A''_k.$$

Now assume that there exist $Q, Q' \in \Delta\mathfrak{A}^\cup$ with

$$Q \restriction \mathfrak{A}^\cap = Q' \restriction \mathfrak{A}^\cap, \tag{59}$$

and $Q \in J$, but $Q' \notin J$. Furthermore, assume that Q, Q' are minimal with these properties in the sense that the number of atoms of $\mathfrak{A}^\cup$ to which Q and Q' assign different probabilities is minimal.

From $Q \neq Q'$ and (59) it follows that there exists an atom C of $\mathfrak{A}^\cap$, and atoms $B_{hk}, B_{h'k'} \subset C$ of $\mathfrak{A}^\cup$, such that

$$\begin{aligned} Q(B_{hk}) &= Q'(B_{hk}) + r \\ Q(B_{h'k'}) &= Q'(B_{h'k'}) - s \end{aligned}$$

for some $r, s > 0$. Assume that $r \leq s$ (the argument for the case $s < r$ proceeds similarly). We show that there exists a sequence

$$(i_0, j_0), (i_1, j_1), \ldots, (i_n, j_n) \tag{60}$$

in $\{1, \ldots, p\} \times \{1, \ldots, q\}$ such that

$$(i_0, j_0) = (h, k), \quad (i_n, j_n) = (h', k'), \tag{61}$$

and for all $h = 1, \ldots, n$:

$$i_h = i_{h-1} \;\text{ or }\; j_h = j_{h-1}, \;\text{ and } B_{i_h, j_h} \neq \emptyset. \tag{62}$$

Once we have such a sequence, we derive a contradiction to the minimality assumption for Q, Q' as follows: we construct a sequence

$$Q = Q_0, Q_1, \ldots, Q_n$$

by defining for all atoms B of $\mathfrak{A}^\cup$ and for $h = 1, \ldots, n$:

$$Q_h(B) := \begin{cases} Q_{h-1}(B) & B \notin \{B_{i_{h-1} j_{h-1}}, B_{i_h j_h}\} \\ Q_{h-1}(B) - r & B = B_{i_{h-1} j_{h-1}} \\ Q_{h-1}(B) + r & B = B_{i_h j_h} \end{cases}$$

(i.e. we just "shift" probability mass r from B_{hk} to $B_{h'k'}$ via the $B_{i_h j_h}$). For all $h = 1, \ldots, n$ then $Q_h \in J$, because $Q_0 \in J$, and $Q_h \upharpoonright \mathfrak{A}' = Q_{h-1} \upharpoonright \mathfrak{A}'$ (if $i_h = i_{h-1}$), or $Q_h \upharpoonright \mathfrak{A}'' = Q_{h-1} \upharpoonright \mathfrak{A}''$ (if $j_h = j_{h-1}$). Thus, $Q_n \in J$, $Q_n \upharpoonright \mathfrak{A}^\cap = Q' \upharpoonright \mathfrak{A}^\cap$, and Q_n agrees with Q' on one atom more than does Q, a contradiction.

It remains to show the existence of the sequence (60). For this we define a relation $(h,k) \rightarrow \cdot$ on $\{1, \ldots, p\} \times \{1, \ldots, q\}$ by: $(h,k) \rightarrow (i,j)$ iff there exists a sequence (60) with $(i_0, j_0) = (h,k)$ and $(i_n, j_n) = (i,j)$ so that (62) holds. Now consider

$$A := \bigcup_{(i,j):(h,k)\rightarrow(i,j)} B_{ij}.$$

As $(h,k) \rightarrow (i,j)$ and $B_{i'j} \neq \emptyset$ implies $(h,k) \rightarrow (i',j)$ (respectively $B_{ij'} \neq \emptyset$ implies $(h,k) \rightarrow (i,j')$), we obtain

$$A = \bigcup_{i:\, \exists j (h,k)\rightarrow(i,j)} A'_i = \bigcup_{j:\, \exists i (h,k)\rightarrow(i,j)} A''_j,$$

which means that $A \in \mathfrak{A}^\cap$ (in fact, $A = C$). From $A \in \mathfrak{A}^\cap$, $B_{hk} \subseteq A$, and $B_{h'k'}$ belonging to the same atom of $\mathfrak{A}^\cap$ as B_{hk}, it follows that $B_{h'k'} \subseteq A$, i.e. $(h,k) \rightarrow (h'k')$. □

From lemmas 3.3 and 3.4 it follows that the set $\Pi_J(P)$ defined in (57) does not depend on the choice of $\mathfrak{A}'$: by lemma 3.4 there exists a unique smallest algebra $\mathfrak{A}^*$ by constraints on which J is defined, and by lemma 3.3 we have for every $\mathfrak{A}' \supseteq \mathfrak{A}^*$:

$$\{\mathscr{J}(Q', P, \mathfrak{A}) \mid Q' \in \Pi_{J \upharpoonright \mathfrak{A}'}(P \upharpoonright \mathfrak{A}')\} = \{\mathscr{J}(Q^*, P, \mathfrak{A}) \mid Q^* \in \Pi_{J \upharpoonright \mathfrak{A}^*}(P \upharpoonright \mathfrak{A}^*)\}.$$

Definition 3.5 Let $\mathfrak{A}$ be an algebra over M, $\mathfrak{A}'$ a finite subalgebra of $\mathfrak{A}$. Let $J \subseteq \Delta_{\mathbb{F}}\mathfrak{A}$ be defined by constraints on $\mathfrak{A}'$, and $P \in \Delta_{\mathbb{F}}\mathfrak{A}$. The set $\Pi_{J \upharpoonright \mathfrak{A}'}(P \upharpoonright \mathfrak{A}')$ is defined by (43). The cross-entropy projection of P onto J then is defined by (57).

We are now ready to define the preferred model semantics for L_p. Recall that it is our aim to identify those models $\mathfrak{M}$ of a L_p-formula ϕ for which the subjective probability measure $Q_{\mathbf{e}}$ minimizes cross-entropy with respect to the statistical measure $P_{|\mathbf{e}|}$, and that this minimization is to be effected only by choosing suitable $Q_{\mathbf{e}}$ for every possible given $P_{|\mathbf{e}|}$, not by selecting any preferred $P_{|\mathbf{e}|}$.

For a probabilistic structure $\mathfrak{M} = (M, \ldots, \mathfrak{F}, \ldots, Q_{\mathbf{e}})$ and $Q \in \Delta_{\mathbb{F}}\mathfrak{A}_{|\mathbf{e}|}$ we denote by $\mathfrak{M}[Q_{\mathbf{e}}/Q]$ the structure $\mathfrak{M}'$ that is obtained by replacing $Q_{\mathbf{e}}$ with Q. For a sufficient probabilistic structure $\mathfrak{M}$, and an L_p-sentence ϕ we define

$$\Delta_{\mathbb{F}}(\phi, \mathfrak{M}) := \{Q \in \Delta_{\mathbb{F}}\mathfrak{A}_{|\mathbf{e}|} \mid \mathfrak{M}[Q_{\mathbf{e}}/Q] \models \phi\}. \tag{63}$$

Thus, $\Delta_{\mathbb{F}}(\phi, \mathfrak{M})$ is the set of subjective probability measures that will turn the non-subjective part $(M, I, \mathfrak{F}, (\mathfrak{A}_n, P_n)_{n \in \mathbb{N}})$ of $\mathfrak{M}$ into a model of ϕ (it is not difficult to show that such a substitution cannot destroy sufficiency).

The following lemma is the main reason for the syntactic restrictions that were imposed on subjective probability terms.

Lemma 3.6 For all $\mathfrak{M}$ and ϕ: $\Delta_{\mathbb{F}}(\phi, \mathfrak{M})$ is defined by constraints on a finite subalgebra $\mathfrak{A}'$ of $\mathfrak{A}_{|\mathbf{e}|}$.

Proof:ϕ contains a finite number of subjective probability terms $\mathrm{prob}(\psi_1(\mathbf{e})), \ldots$ $\ldots, \mathrm{prob}(\psi_k(\mathbf{e}))$. Membership of $Q \in \Delta\mathfrak{A}_{|\mathbf{e}|}$ in $\Delta(\phi, \mathfrak{M})$ only depends on the values $Q((\mathfrak{M}, \boldsymbol{v})(\psi_i(\boldsymbol{v}))$ $(i = 1, \ldots, k)$. By the condition that the ψ_i do not contain any occurrences of $\mathrm{prob}(\cdot)$, the sets $(\mathfrak{M}, \boldsymbol{v})(\psi_i(\boldsymbol{v}))$ do not depend on the component $Q_{\mathbf{e}}$ of $\mathfrak{M}$. Let $\mathfrak{A}'$ be the finite subalgebra of $\mathfrak{A}_{|\mathbf{e}|}$ generated by the sets $(\mathfrak{M}, \boldsymbol{v})(\psi_i(\boldsymbol{v}))$. Then $\mathfrak{A}'$ is a finite algebra so that for every $Q \in \Delta\mathfrak{A}_{|\mathbf{e}|}$ the validity of $\mathfrak{M}[Q_{\mathbf{e}}/Q] \models \phi$ is determined by the values of Q on $\mathfrak{A}'$. □

No analogue of lemma 3.6 would hold if we dropped either the prohibition of nested subjective probability terms, or of free variables in subjective probability terms. Together, definition 3.5 and lemma 3.6 permit the following final definition of the inductive entailment relation $\models\!\!\!\approx$ for $\mathscr{L}_{ip}$.

Definition 3.7 Let $\phi \in L_p(S, \mathbf{e})$, $\mathfrak{M} = (M, \ldots, Q_{\mathbf{e}})$ a sufficient probabilistic structure for $(S, \mathbf{e})$. $\mathfrak{M}$ is called a *preferred model* of ϕ, written $\mathfrak{M} \models\!\!\!\approx \phi$, iff

$$Q_{\mathbf{e}} \in \Pi_{\Delta_{\mathbb{F}}(\phi, \mathfrak{M})}(P_{|\mathbf{e}|}). \tag{64}$$

For $\phi, \psi \in L_p(S, \mathbf{e})$ we define: $\phi \models\!\!\!\approx \psi$ iff $\mathfrak{M} \models \psi$ for every preferred model $\mathfrak{M}$ of ϕ.

3.3 Inductive Reasoning in $\mathscr{L}_{ip}$

Having formally defined our inductive entailment relation $\models\!\!\!\approx$, we now investigate some of its logical properties. Our first goal is to verify that the relation $\models\!\!\!\approx$ indeed supports the patterns of inductive inference described in sections 1.1 and 3.1, which motivated the approach we have taken. This is established in the following using the structural properties of *CE*-projections described in theorems A.5 (system independence) and A.6 (subset independence).

At the very outset we stipulated that the relation $\models\!\!\!\approx$ should implement direct inference, where direct inference is applicable. From corollary A.7 one immediately obtains that the inductive inference

$$[\psi(\boldsymbol{v})]_{\boldsymbol{v}} > 0 \wedge [\phi(\boldsymbol{v}) \mid \psi(\boldsymbol{v})]_{\boldsymbol{v}} = r \wedge \mathrm{prob}(\psi[\mathrm{e}]) = 1 \models\!\!\!\approx \mathrm{prob}(\phi[\mathrm{e}]) = r \tag{65}$$

is valid in $\mathscr{L}_{ip}$ for all formulas ϕ, ψ. Usually, however, our total knowledge does not have the form of the premise of (65): one does not only know that $\psi[\mathrm{e}]$ is true for a single property ψ, but rather that $\psi_1[\mathrm{e}], \ldots, \psi_n[\mathrm{e}]$ are true. Assuming the necessary statistical knowledge as well, our premise then is

$$\wedge_{i=1}^{n}([\psi_i(\boldsymbol{v})]_{\boldsymbol{v}} > 0 \wedge [\phi(\boldsymbol{v}) \mid \psi_i(\boldsymbol{v})]_{\boldsymbol{v}} = r_i \wedge \mathrm{prob}(\psi_i[\mathrm{e}]) = 1). \tag{66}$$

The question of what to inductively infer from this body of knowledge is essentially the problem of the choice of the best *reference class* for direct inference (Pollock

1983, Kyburg 1983). The original prescription by Reichenbach (1949) was to take the smallest reference class for which reliable statistics exist. We cannot follow this principle in $\mathscr{L}_{ip}$, because, first, in our framework we do not have the means to distinguish the reliabilities of two statistical statements $[\phi(\boldsymbol{v}) \mid \psi_i(\boldsymbol{v})]_{\boldsymbol{v}} = r_i$ and $[\phi(\boldsymbol{v}) \mid \psi_k(\boldsymbol{v})]_{\boldsymbol{v}} = r_k$, and second, from the logical equivalence of (66) and

$$\wedge_{i=1}^{n}([\psi_i(\boldsymbol{v})]_{\boldsymbol{v}} > 0 \wedge [\phi(\boldsymbol{v}) \mid \psi_i(\boldsymbol{v})]_{\boldsymbol{v}} = r_i) \wedge \mathrm{prob}(\wedge_{i=1}^{n}\psi_i[\mathbf{e}]) = 1, \tag{67}$$

it follows with (65) that from (66) we will always have to infer

$$[\wedge_{i=1}^{n}\psi_i(\boldsymbol{v})]_{\boldsymbol{v}} > 0 \rightarrow \mathrm{prob}(\phi[\mathbf{e}]) = [\phi(\boldsymbol{v}) \mid \wedge_{i=1}^{n}\psi_i(\boldsymbol{v})]_{\boldsymbol{v}}. \tag{68}$$

Thus, we always base direct inference on the smallest reference class that $\mathbf{e}$ belongs to, whether or not the statistics for this reference class can be deemed reliable – or even are available. In extreme cases this leads to inferences that may seem overly conservative: consider

$$\begin{aligned} \phi_1 \equiv \;& [\mathtt{IIA}(d) \mid \neg\mathtt{Drinks}(d)]_d = 0.01 \\ & \wedge \mathrm{prob}(\neg\mathtt{Drinks}(\mathit{jones})) = 1, \\ \phi_2 \equiv \;& [\mathtt{IIA}(d) \mid \mathtt{Drives}(\mathtt{Toyota}, d)]_d = 0.01 \\ & \wedge \mathrm{prob}(\mathtt{Drives}(\mathtt{Toyota}, \mathit{jones})) = 1. \end{aligned}$$

Then $\phi_1 \mid\!\approx \mathrm{prob}(\mathtt{IIA}(\mathit{jones})) = 0.01$, and $\phi_2 \mid\!\approx \mathrm{prob}(\mathtt{IIA}(\mathit{jones})) = 0.01$, but not

$$\phi_1 \wedge \phi_2 \mid\!\approx \mathrm{prob}(\mathtt{IIA}(\mathit{jones})) = 0.01. \tag{69}$$

This is because we will infer

$$\phi_1 \wedge \phi_2 \mid\!\approx \mathrm{prob}(\mathtt{IIA}(\mathit{jones})) = [\mathtt{IIA}(d) \mid \neg\mathtt{Drinks}(d) \wedge \mathtt{Drives}(\mathtt{Toyota}, d)]_d. \tag{70}$$

Going from (70) to (69) amounts to an implicit default inference about statistical probabilities

$$\begin{aligned} & [\mathtt{IIA}(d) \mid \neg\mathtt{Drinks}(d)]_d = 0.01 \wedge [\mathtt{IIA}(d) \mid \mathtt{Drives}(\mathtt{Toyota}, d)]_d = 0.01 \\ & \qquad\qquad \mid\!\approx \; [\mathtt{IIA}(d) \mid \neg\mathtt{Drinks}(d) \wedge \mathtt{Drives}(\mathtt{Toyota}, d)]_d = 0.01, \end{aligned}$$

which $\mathscr{L}_{ip}$ is not designed to do.

Basing direct inference on the narrowest possible reference class can lead to difficulties when the subject of the direct inference ($\mathbf{e}$ in our case) is referenced in the definition of the reference class (see e.g. (Pollock 1983, Section 6)). In particular, one then might consider the single point reference class $\{\mathbf{e}\}$. and argue that direct inference in $\mathscr{L}_{ip}$ must always identify $\mathrm{prob}(\phi(\mathbf{e}))$ with $[\phi(\boldsymbol{v}) \mid \boldsymbol{v} = \mathbf{e}]_{\boldsymbol{v}}$. Since this statistical probability can only assume the values 0 or 1 (according to whether $\phi(\mathbf{e})$ holds), it might therefore appear as though

$$\mathrm{prob}(\phi(\mathbf{e})) = 0 \vee \mathrm{prob}(\phi(\mathbf{e})) = 1. \tag{71}$$

is valid in $\mathscr{L}_{ip}$ with respect to $\approx\!\!\!\!\models$ -entailment. As in the derivation of (24), however, this argument is based on incorrectly using $\mathbf{e}$ in the expression $[\phi(\boldsymbol{v}) \mid \boldsymbol{v} = \mathbf{e}]_{\boldsymbol{v}}$ like a standard constant symbol. The syntactic condition that $\mathbf{e}$ must always appear within the scope of a prob()-operator prevents the construction of reference classes involving $\mathbf{e}$.

When our knowledge base is of a form that makes Jeffrey's rule applicable, then we derive from corollary A.7 that $\models\!\approx$ coincides with Jeffrey's rule.

Leaving the elementary cases of direct inference and Jeffrey's rule behind, we next consider some logical properties of $\mathscr{L}_{ip}$ that in a more general way reflect the system- and subset-independence properties of *CE*-projections. First, we use system-independence to derive the general (logical) independence property of $\models\!\approx$, an instance of which was illustrated by (44) and (45).

Theorem 3.8 Let S be a vocabulary, $\mathbf{e}$ and $\mathbf{f}$ two disjoint tuples of event symbols. Let $\phi_{\mathbf{e}}, \psi_{\mathbf{e}}(\boldsymbol{v}) \in L_p(S, \mathbf{e})$, $\phi_{\mathbf{f}}, \psi_{\mathbf{f}}(\boldsymbol{w}) \in L_p(S, \mathbf{f})$, with $|\,\boldsymbol{v}\,| = |\,\mathbf{e}\,|$ and $|\,\boldsymbol{w}\,| = |\,\mathbf{f}\,|$. Then

$$\phi_{\mathbf{e}} \wedge \phi_{\mathbf{f}} \models\!\approx \mathrm{prob}(\psi_{\mathbf{e}}[\mathbf{e}] \wedge \psi_{\mathbf{f}}(\mathbf{f})) = \mathrm{prob}(\psi_{\mathbf{e}}[\mathbf{e}])\mathrm{prob}(\psi_{\mathbf{f}}(\mathbf{f})). \tag{72}$$

Proof: Consider a probabilistic structure $\mathfrak{M}$ for $(S, (\mathbf{e}, \mathbf{f}))$. The set $\Delta(\phi_{\mathbf{e}} \wedge \phi_{\mathbf{f}}, \mathfrak{M})$ is defined by constraints on a finite algebra $\mathfrak{A}^{\times} = \mathfrak{A} \times \mathfrak{A}' \subset \mathfrak{A}_{|\,\mathbf{e},\,\mathbf{f}\,|}$, and its restriction $J^{\times}$ to $\mathfrak{A}^{\times}$ has the form

$$\{Q \in \Delta\mathfrak{A}^{\times} \mid Q \restriction \mathfrak{A} \in J_{\mathbf{e}}, Q \restriction \mathfrak{A}' \in J_{\mathbf{f}}\}$$

for $J_{\mathbf{e}} \subseteq \Delta\mathfrak{A}, J_{\mathbf{f}} \subseteq \Delta\mathfrak{A}'$. The restriction $P^{\times}$ of the statistical distribution $P_{|\,\mathbf{e},\,\mathbf{f}\,|}$ to $\mathfrak{A}^{\times}$ is a product measure, so that every

$$Q \in \Pi_{J^{\times}}(P^{\times})$$

also is a product measure on $\mathfrak{A}^{\times}$. The theorem now follows from theorem A.5, and by observing (using lemma 3.3) that the Jeffrey-extension $\mathscr{J}(Q, P_{|\,\mathrm{e},\,\mathrm{f}\,|}, \mathfrak{A}_{|\,\mathrm{e},\,\mathrm{f}\,|})$ preserves the product property for sets of the form $A \times B$ with $A \in \mathfrak{A}_{|\,\mathbf{e}\,|}$, $B \in \mathfrak{A}_{|\,\mathbf{f}\,|}$.
□

The next theorem transforms subset-independence (theorem A.6) into a statement about the coherency of conditional reasoning in $\mathscr{L}_{ip}$.

Theorem 3.9 Let $\phi_{|\gamma}, \psi_{|\gamma} \in L_p$ only contain subjective probability terms of the form $\mathrm{prob}(\phi[\mathrm{e}] \mid \gamma[\mathrm{e}])$ for some fixed $\gamma \in L_p$. Let ϕ, ψ be the sentences obtained from $\phi_{|\gamma}, \psi_{|\gamma}$ by replacing each term $\mathrm{prob}(\phi[\mathrm{e}] \mid \gamma[\mathrm{e}])$ with the corresponding unconditional term $\mathrm{prob}(\phi[\mathrm{e}])$. Then

$$\phi_{|\gamma} \wedge \mathrm{prob}(\gamma[\mathbf{e}]) > 0 \models\!\approx \psi_{|\gamma} \tag{73}$$

iff

$$\phi \wedge \mathrm{prob}(\gamma[\mathbf{e}]) = 1 \models\!\approx \psi. \tag{74}$$

Note that adding the conjunct $\mathrm{prob}(\gamma[\mathbf{e}]) > 0$ to the premise of (73) means that there is no ambiguity in the interpretations of the conditional probability terms in $\phi_{|\gamma}$ and $\psi_{|\gamma}$, so that the theorem holds independent from the conventions adopted for dealing with conditioning events of probability zero. The proof of the theorem is similar to that of the previous one, by first noting that the structure of the set $\Delta(\phi_{|\gamma} \wedge \mathrm{prob}(\gamma[\mathbf{e}]) > 0, \mathfrak{M})$ is a special case of the form described in theorem A.6, then applying that theorem, and finally observing that the structural property expressed in (106) is preserved under Jeffrey extensions.

In section 1.1 we said that $\mathscr{L}_{ip}$ is not intended to model any inductive inferences about statistical probabilities, based on (even numerous) single case observations. By defining preferred models in terms of the condition (64) on the subjective distribution $Q_{\mathbf{e}}$ for any given statistical distribution $P_{|\mathbf{e}|}$ this goal is essentially realized, but with the following caveat: statistical distributions $P_{|\mathbf{e}|}$ for which $\Pi_{\Delta_{\mathbb{F}}(\phi,\mathfrak{M})}(P_{|\mathbf{e}|})$ is empty are ruled out. This means, in particular, that distributions $P_{|\mathbf{e}|}$ are ruled out for which $\Delta_{\mathbb{F}}(\phi, \mathfrak{M})$ does not contain any $Q_{\mathbf{e}}$ with $Q_{\mathbf{e}} \ll P_{|\mathbf{e}|}$ (cf. (43) and definition 3.7). In consequence, for example the following is a valid inference pattern in $\mathscr{L}_{ip}$:

$$\mathrm{prob}(\phi(\mathbf{e})) > 0 \;\;|\!\!\approx\;\; [\phi(\boldsymbol{v})]_{\boldsymbol{v}} > 0. \tag{75}$$

While, in principle, this is a default inference about statistical probabilities from subjective probabilities, (75) may still be considered unproblematic even from our conservative point of view, because it just amounts to the reasonable constraint that in preferred models we cannot assign nonzero probabilities to events $\mathbf{e}$ having some statistically impossible property ϕ. Observe that (75) means that for $|\!\!\approx$ we obtain a strengthening of (29).

The set $\Pi_{\Delta_{\mathbb{F}}(\phi,\mathfrak{M})}(P_{|\mathbf{e}|})$ can also be empty because the infimum is not attained in *CE*-minimization. Consider, for example, the sentence

$$\phi_{76} = \;\; ([\psi(v)]_v = 0.3 \vee [\psi(v)]_v = 0.5) \wedge \mathrm{prob}(\psi(\mathrm{e})) > 0.4. \tag{76}$$

For any model $\mathfrak{M}$ of ϕ_{76} with $P_{\mathrm{e}}((\mathfrak{M}, v)(\psi)) = 0.3$ then $\Pi_{\Delta_{\mathbb{F}}(\phi,\mathfrak{M})}(P_{|\mathbf{e}|}) = \emptyset$, because $CE(\cdot, P_{\mathrm{e}})$ is not minimized over the open interval $]0.4, 1]$ defining $\Delta_{\mathbb{F}}(\psi, \mathfrak{M})$. When $P_{\mathrm{e}}((\mathfrak{M}, v)(\psi)) = 0.5$, on the other hand, the infimum is attained for $Q \in \Delta_{\mathbb{F}}(\psi, \mathfrak{M})$ with $Q((\mathfrak{M}, v)(\psi)) = 0.5$. Thus, ϕ_{76} only has preferred models in which the statistical probability of ψ is 0.5, i.e.

$$\phi_{76} \;|\!\!\approx [\psi(v)]_v = 0.5.$$

Thus, some potentially undesired inferences can occur when constraints on the subjective distribution define non-closed sets $\Delta_{\mathbb{F}}(\phi, \mathfrak{M})$. This is a typical limitation of methods based on minimizing distance measures, and often circumvented by prohibiting non-closed constraint sets. In the very general language L_p it is difficult to enforce closedness of $\Delta_{\mathbb{F}}(\phi, \mathfrak{M})$ by a simple syntactic condition on ϕ. Such a condition, therefore, has not been imposed in the basic definitions. However, in practical modeling with L_p some attention should be paid to the question whether the sets $\Delta_{\mathbb{F}}(\phi, \mathfrak{M})$ will be closed (see also section 4.2).

3.4 Axiomatization

In this section we obtain a completeness result for the inductive entailment relation $\mid\!\approx$. The result is derived by showing that for a given L_p-sentence ϕ there exists a recursively enumerable set $\mathrm{MinCE}(\phi) \subseteq L_p$ that axiomatizes inductive entailment, i.e.

$$\phi \mid\!\approx \psi \quad \text{iff} \quad \mathrm{MinCE}(\phi) \models \psi \qquad\qquad (\psi \in L_p). \tag{77}$$

By the completeness result for strict inference we then obtain a completeness result for $\mid\!\approx$. This approach of capturing the preferred models of ϕ by adjoining to ϕ a set of axioms dependent on ϕ is closely related to the *circumscription* framework (McCarthy 1980) in nonmonotonic reasoning.

To establish (77) it is sufficient to find a set $\mathrm{MinCE}(\phi)$ that axiomatizes the class of preferred models of ϕ up to elementary equivalence, i.e. to obtain that a probabilistic structure $\mathfrak{M}$ is a model of $\mathrm{MinCE}(\phi)$ iff it is elementarily equivalent to a structure $\mathfrak{M}'$ with $\mathfrak{M}' \mid\!\approx \phi$ (recall that two structures are called elementarily equivalent iff they satisfy the same sentences). For a structure $\mathfrak{M} = (\ldots, (\mathfrak{A}_n, P_n)_{n\in\mathbb{N}}, Q_{\mathbf{e}})$ to be a preferred model of ϕ, by definition, is equivalent for $\mathfrak{M}$ to satisfy the condition

$$Q_{\mathbf{e}} \in \Pi_{\Delta_{\mathbb{F}}(\phi,\mathfrak{M})}(P_{|\mathbf{e}|}). \tag{78}$$

Elementary equivalence to a preferred model, on the other hand, is guaranteed by the weaker condition

$$Q_{\mathbf{e}} \restriction \mathfrak{A}^* \in \Pi_{\Delta_{\mathbb{F}}(\phi,\mathfrak{M})\restriction\mathfrak{A}^*}(P_{|\mathbf{e}|} \restriction \mathfrak{A}^*), \tag{79}$$

where $\mathfrak{A}^* \subseteq \mathfrak{A}_{|\mathbf{e}|}$ is the subalgebra consisting of those sets that are definable by an L_p-formula without parameters, i.e. $A \in \mathfrak{A}^*$ iff there exists $\psi(\boldsymbol{v}) \in L_p$ with $A = (\mathfrak{M}, \boldsymbol{v})(\psi)$. That (79) implies elementary equivalence to a preferred model follows from the fact that any two structures $\mathfrak{M}$ and $\mathfrak{M}'$ that differ only with respect to $Q_{\mathbf{e}}$-values for elements $A \in \mathfrak{A}_{|\mathbf{e}|} \setminus \mathfrak{A}^*$ are elementarily equivalent, and that any structure $\mathfrak{M}$ that satisfies (79) can be modified into a preferred model of ϕ by only changing $Q_{\mathbf{e}}$-values on $\mathfrak{A}_{|\mathbf{e}|} \setminus \mathfrak{A}^*$. Thus, it will be sufficient to capture with $\mathrm{MinCE}(\phi)$ the class of models that satisfy (79).

Using that we have defined *CE*-projections on infinite algebras via the two steps (43) and (57), we can split (79) into two parts: abbreviating $\Delta_{\mathbb{F}}(\phi, \mathfrak{M})$ by J, and letting $\mathfrak{A}'$ be a finite subalgebra by constraints on which J is defined, we obtain out of (43) the condition

$$Q_{\mathbf{e}} \restriction \mathfrak{A}' \in \Pi_{J\restriction\mathfrak{A}'}(P_{|\mathbf{e}|} \restriction \mathfrak{A}'). \tag{80}$$

When (80) is fulfilled, and $A_1, \ldots, A_L$ are the atoms of $\mathfrak{A}'$, then the defining equation (57) can be expressed by

$$Q_{\mathbf{e}}(B) = \sum_{\substack{h=1 \\ P_{|\mathbf{e}|}(A_h)>0}}^{L} Q_{\mathbf{e}}(A_h) P_{|\mathbf{e}|}(B \mid A_h) \qquad (B \in \mathfrak{A}^*). \tag{81}$$

We now axiomatize (80) by a single L_p-formula, and (81) by a schema, ranging over the B. Our first task is to identify a suitable algebra $\mathfrak{A}'$, and its atoms $A_1, \ldots, A_L$. As in the proof of lemma 3.6 let

$$\mathrm{prob}(\psi_1[\mathbf{e}]), \ldots, \mathrm{prob}(\psi_n[\mathbf{e}])$$

be the subjective probability terms contained in ϕ. Then $\Delta_{\mathbb{F}}(\phi, \mathfrak{M})$ is defined by constraints on the algebra $\mathfrak{A}'$ generated by the extensions of the ψ_i. The atoms of $\mathfrak{A}'$ are the nonempty extensions of the formulas

$$\alpha_j(\boldsymbol{v}) := \wedge_{i=1}^{n} \tilde{\psi}_i(\boldsymbol{v}) \quad (\tilde{\psi}_i(\boldsymbol{v}) \in \{\psi_i(\boldsymbol{v}), \neg\psi_i(\boldsymbol{v})\},\ j = 1, \ldots, 2^n).$$

As a first building block for the formalization of (80) we can now formulate an L_p-formula that defines as a subset of $\mathbb{F}^{2^n}$ the set of all probability measures on $\mathfrak{A}'$:

$$\begin{aligned}\delta(x_1, \ldots, x_{2^n}) :\equiv\ & \bigwedge_{j=1}^{2^n} x_j \geq 0 \wedge \sum_{j=1}^{2^n} x_j = 1 \\ & \wedge \bigwedge_{j=1}^{2^n} (\neg\exists \boldsymbol{v} \alpha_j(\boldsymbol{v}) \rightarrow x_j = 0).\end{aligned}$$

Now let $\phi[\mathrm{prob}/\boldsymbol{x}]$ denote the formula that is obtained from ϕ by substituting for every term $\mathrm{prob}(\psi_i[\mathbf{e}])$ the term $x_{j_1} + \ldots + x_{j_k}$ where $k = 2^{n-1}$, and $\{j_1, \ldots, j_k\} \subset \{1, \ldots, 2^n\}$ is the collection of indices j_h for which the atom α_{j_h} is contained in ψ_i (i.e. α_{j_h} is a conjunction in which ψ_i appears un-negated). For the formula

$$\iota(\boldsymbol{x}) := \delta(\boldsymbol{x}) \wedge \phi[\mathrm{prob}/\boldsymbol{x}] \tag{82}$$

and a probabilistic structure $\mathfrak{M}$ we then have

$$(\mathfrak{M}, \boldsymbol{x})(\iota(\boldsymbol{x})) = \Delta_{\mathbb{F}}(\phi, \mathfrak{M}) \upharpoonright \mathfrak{A}'. \tag{83}$$

The formula

$$\zeta(\boldsymbol{x}) :\equiv \bigwedge_{j=1}^{2^n} ([\alpha_j(\boldsymbol{v})]_{\boldsymbol{v}} = 0 \rightarrow x_j = 0) \tag{84}$$

encodes the condition of absolute continuity with respect to the statistical distribution on the algebra $\mathfrak{A}'$. In particular, the sentence

$$\zeta[\mathrm{prob}] :\equiv \bigwedge_{j=1}^{2^n} ([\alpha_j(\boldsymbol{v})]_{\boldsymbol{v}} = 0 \rightarrow \mathrm{prob}(\alpha_j[\mathbf{e}]) = 0) \tag{85}$$

says that $Q_{\mathbf{e}} \upharpoonright \mathfrak{A}' \ll P_{|\mathbf{e}|} \upharpoonright \mathfrak{A}'$. We now can axiomatize (80) by the L_p-sentence

$$\begin{aligned}&\zeta[\mathrm{prob}] \wedge \forall \boldsymbol{x}((\iota(\boldsymbol{x}) \wedge \zeta(\boldsymbol{x})) \rightarrow \\ &\qquad \sum_{j:[\alpha_j(\boldsymbol{v})]_{\boldsymbol{v}}>0} x_j Log \frac{x_j}{[\alpha_j(\boldsymbol{v})]_{\boldsymbol{v}}} \geq \sum_{j:[\alpha_j(\boldsymbol{v})]_{\boldsymbol{v}}>0} \mathrm{prob}(\alpha_j[\mathbf{e}]) Log \frac{\mathrm{prob}(\alpha_j[\mathbf{e}])}{[\alpha_j(\boldsymbol{v})]_{\boldsymbol{v}}})\end{aligned} \tag{86}$$

(we are here taking some liberties with the syntax of L_p, but one can easily expand this formula so as to eliminate the sum-expressions, and obtain a proper L_p-sentence).

To encode (81), let B be defined by the formula $\beta(\boldsymbol{v})$. Then (81) can be written in L_p as

$$\mathrm{prob}(\beta[\mathbf{e}]) = \sum_{j:[\alpha_j(\boldsymbol{v})]_{\boldsymbol{v}}>0} \mathrm{prob}(\alpha_j[\mathbf{e}])[\beta(\boldsymbol{v}) \mid \alpha_j(\boldsymbol{v})]_{\boldsymbol{v}}. \tag{87}$$

Taking the union over all L_p-formulas $\beta(\boldsymbol{v})$ with $|\boldsymbol{v}| = |\mathbf{e}|$ turns (87) into a recursively enumerable sentence schema.

Finally, let MinCE(ϕ) consist of ϕ, of (86), and all instances of (87). Clearly there exists an algorithm that for any given sentence ϕ enumerates MinCE(ϕ) (we only need a uniform method to generate the atoms α_j determined by ϕ, and then simply list(86) and all instances of (87)). Also, by our derivation of MinCE(ϕ), clearly (77) is satisfied. Thus, the enumeration algorithm for MinCE(ϕ), together with a complete inference system for $\models$, constitutes a complete inference system for $\approx\!\!\!\!|\;$.

4 Related Work and Conclusion

4.1 Related Work

Closely related to our logic of strict inference, $\mathscr{L}_p$, are the probabilistic first-order logics of Bacchus (1990*b*, 1990*a*) and Halpern (1990). Our logic of inductive inference, $\mathscr{L}_{ip}$, on the other hand, has to be compared with the random worlds method of Bacchus, Grove, Halpern, and Koller (Bacchus et al. (1992, 1997), Grove et al. (1992*a*, 1992*b*)).

There are two main differences between our logic $\mathscr{L}_p$ and the combined subjective and statistical probability logic $\mathscr{L}_3$ of Halpern (1990). The first difference lies in basing the semantics of $\mathscr{L}_p$ on arbitrary lrc-field valued measures, whereas the semantics of $\mathscr{L}_3$ is based on real-discrete measures alone. As a result, a completeness result corresponding to our theorem 2.12 cannot be obtained for $\mathscr{L}_3$ (Abadi & J.Y.Halpern 1994). However, measures in more general algebraic structures were also already used by Bacchus (1990*a*) to obtain a completeness result for his statistical probability logic, and the same approach could clearly also be directly applied to Halpern's $\mathscr{L}_3$. The second difference between $\mathscr{L}_p$ and $\mathscr{L}_3$, therefore, is the much more significant one: in $\mathscr{L}_3$ statistical and subjective probabilities are interpreted by probability measures on the domains of first-order structures, and probability measures on sets of such structures (or possible worlds), respectively (leading to *type-3 probability structures*). As a result, the logic does not enforce any connections between statistical and subjective probabilities, or, more generally, domain knowledge and subjective probabilities. For example, the sentence

$$\neg\exists v\phi(v) \wedge \mathrm{prob}(\phi(\mathrm{e})) = 0.5 \tag{88}$$

is satisfiable in $\mathscr{L}_3$ by a type-3 structure containing a possible world that does not have any elements with property ϕ, and also containing possible worlds in which $\phi(\mathrm{e})$ is true (when interpreting (88) as a sentence in $\mathscr{L}_3$, the symbol e is considered as a standard constant). Halpern (1990) also shows that some dependencies between statistical and subjective probabilities are obtained in $\mathscr{L}_3$ when the semantics is restricted to type-3 structures in which all relation and function symbols are *rigid*, i.e. have the same interpretation in all possible worlds, and only the interpretations of some constant symbols are allowed to vary over the possible worlds. These dependencies are very weak, however, and do "not begin to settle the issue of how to connect statistical information with degrees of belief" (Halpern 1990). Our probabilistic structures are closely related to these rigid type-3 structures. In fact, we can view a probabilistic structure in our sense as a superimposition of the possible worlds in a rigid type-3 structure, where non-rigid constant symbols now become our event symbols, and the distribution $Q_{\mathbf{e}}$ represents their distribution in the different possible worlds. This collapsing of the possible worlds into a single structure gives us the crucial technical advantage that subjective and statistical probabilities are defined on the same space, and their discrepancy can be measured by cross-entropy.

The statistical probability logics of Bacchus and Halpern serve as the foundation for the random-worlds method of Bacchus, Grove, Halpern, and Koller (Bacchus et al. (1992, 1997), Grove et al. (1992*a*, 1992*b*)). Aim of this approach is to assign to pairs ϕ, ψ of formulas in the statistical representation language a degree of belief $\Pr(\phi|\psi)$ in the proposition ϕ, given the knowledge ψ. The definition of $\Pr(\phi|\psi)$ proceeds by considering for fixed $n \in \mathbb{N}$ the fraction $\Pr_n(\phi|\psi)$ of models of ψ over domain $\{1, \ldots, n\}$ that also satisfy ϕ, and to define $\Pr(\phi|\psi) := lim_{n\rightarrow\infty}\Pr_n(\phi|\psi)$, provided that limit exists.

Like our logic $\mathscr{L}_{ip}$, the random worlds method derives much of its motivation from direct inference. A typical example to which the method would be applied is

$$\begin{aligned} \psi &\equiv [\mathtt{IIA}(d)|\neg\mathtt{Drinks}(d)]_d = 0.01 \wedge \neg\mathtt{Drinks}(jones) && (89) \\ \phi &\equiv \mathtt{IIA}(jones), && (90) \end{aligned}$$

for which the random-worlds method yields the direct inference $\Pr(\phi|\psi) = 0.01$. The similarity of motivation, and a connection of the random-worlds method with entropy maximization (Grove, Halpern & Koller 1992*b*), at first sight suggests a fairly close relationship between that method and $\mathscr{L}_{ip}$. On closer examination it turns out, however, that the two frameworks differ substantially with respect to fundamental mathematical properties. The first major difference between the two approaches is that the random-worlds method does not permit to include in the input information ψ any prior constraints on degrees of belief. A second difference lies in the fact that the random-worlds method leads to inferences that go very much beyond the type of inductive probabilistic inferences supported by $\mathscr{L}_{ip}$. In particular, the random-worlds method also leads to default inferences about the statistical distribution, and give, e.g., the degree of belief $\Pr([\mathtt{Drinks}(d)]_d = 0.5|[\mathtt{Drinks}(d)]_d \geq 0.3) = 1$. One sees that, thus, the random-worlds method does not model inductive probabilistic reasoning as we understand it – as an inference pattern that is strictly directed from general (statistical) knowledge to beliefs about a particular case – but leads to a much stronger form of probabilistic default inferences.

Another vital difference arises out of the random-worlds method's commitment to finite domains: if ϕ is a sentence that is not satisfiable on finite domains, and ψ is any sentence, then we obtain $\Pr(\phi|\psi) = 0$; no corresponding phenomenon occurs in $\mathscr{L}_{ip}$. Finally, the random-worlds method differs from $\mathscr{L}_{ip}$ greatly with respect to computational properties. As shown in (Grove, Halpern & Koller 1992*a*), the set of pairs (ϕ, ψ) for which $\Pr(\phi|\psi)$ is defined, i.e. the limit $lim_{n\rightarrow\infty}\Pr_n(\phi|\psi)$ exists, is not recursively enumerable. Thus, there exists no complete proof system for the random-worlds method (a solution to this problem by a move to generalized probabilities here is infeasible, as the very definition of the degrees of belief as limits of sequences of rational numbers is tied to the real number system).

In section 1.1 we argued that for our inductive inference problem a conservative approach is appropriate for combining partial prior information with new information: we simply combine each possible exact prior (i.e. statistical distribution) with the new information (i.e. subjective probability constraints). It is instructive, though, to compare this to some more adventurous rules that have been considered in the literature. A very natural possibility is to perform *CE*-minimization over both the statistical and the subjective probability distribution, i.e. preferred models will be those in which $CE(Q_{\mathbf{e}}, P_{|\mathbf{e}|})$ is minimal for all feasible choices of $Q_{\mathbf{e}}$ and $P_{|\mathbf{e}|}$ (given the non-probabilistic part of the model). This is an instance of *revising based on similarity relationships* (Moral & Wilson 1995). This approach is also closely related to the *classical (maximum likelihood) update rule* of Gilboa and Schmeidler (1993): according to that rule a set C of priors is updated based on the observation of an event A by selecting from C those distributions that assign maximal likelihood to A. If we again identify the categorical observation A with a probability constraint $\text{prob}(A) = 1$, then this means that we select all distributions $q \in C$ with $\min_{p:p(A)=1} CE(p, q) = \min_{q':q' \in C} \min_{p:p(A)=1} CE(p, q)$. Thus, the rule by Gilboa and Schmeidler can also be understood as *CE*-minimization in two arguments (though originally restricted to categorical observations); however, here the result of the updating consists of distributions selected from the set of priors, not from the set defined by the new constraints.

To compare such stronger update rules with our conservative rule, consider the following example:

$$[\psi(v)]_v \geq [\phi(v)]_v \ \wedge \ \text{prob}(\phi(e)) = 1. \tag{91}$$

According to our conservative inference rule, we apply direct inference to every statistical distribution satisfying the statistical constraint in (91). These include for every $q \in [0, 1]$ distributions with $[\psi(v) \mid \phi(v)]_v = q$. Consequently, we will not derive any non-trivial bounds on $\text{prob}(\psi(e))$. If we perform *CE*-minimization in both arguments, then we will effectively only consider statistical distributions with $[\psi(v)]_v = [\phi(v)]_v = 1$, and derive $\text{prob}(\psi(e)) = 1$. This may not seem unreasonable based on the abstract formulation (91), but consider e.g. the case where $\psi(v) = \texttt{Drives}(\texttt{Toyota}, v)$ and $\phi(v) = \texttt{Drives}(\texttt{RollsRoyce}, v)$.

4.2 Conclusion

To formalize the process of inductive probabilistic reasoning within an expressive logical framework we have defined the logic $\mathscr{L}_{ip}$ with its inductive entailment relation $\mathrel{|}\!\approx$. Three design principles have largely guided the definition of $\mathscr{L}_{ip}$: expressiveness, completeness, and epistemic justifiability. Expressiveness: the logic provides a rich first-order representation language that enables the encoding of complex probabilistic information. Completeness: the expressiveness of the language should be complemented with a powerful deductive system. We have obtained a complete deductive system for lrc-field valued probabilities, and have furthermore established a strong agreement between the behaviors of real-valued and lrc-field valued probabilities (especially with regard to cross-entropy minimization). Combined these results entail a strong characterization of the deductive power of our inference system also with respect to real-valued probabilities.

Epistemic justifiability: it was our aim to model with the inductive entailment relation $\approx\!\!\!\!\mid$ only a well-justified pattern of defeasible probabilistic reasoning – how statistical information enables us to refine an already partially formed subjective probability assignment. For this particular inference pattern we argue that cross-entropy minimization relative to every possible statistical distribution is the adequate formal model (more fully than in the present paper this argument is given in (Jaeger 1995*b*) and (Jaeger 1995*a*)). The resulting relation $\mid\!\approx$ is necessarily weak when only little statistical information is available. However, in typical applications one can expect the statistical background information to be much more specific than the partial subjective probability assignments made in the observation of a single event, in which case $\mid\!\approx$ will lead to strong conclusions.

The full logic $\mathscr{L}_{ip}$ should be regarded as a rich reference logic for the theoretical analysis of the formal rules of inductive probabilistic reasoning. For practical applications and implementations one should consider suitable fragments of this logic, e.g. the probabilistic description logics described in (Jaeger 1994*b*). Such fragments can reduce the complexities of reasoning in $\mathscr{L}_{ip}$ in several ways: they can enforce the closure of the sets $\Delta_{\mathbb{F}}(\phi, \mathfrak{M})$, so that some of the difficulties described in section 3.3 are avoided; they can further reduce the discrepancy between real-valued and lrc-field valued probabilities, and thereby become complete also for real-valued probabilities; finally, and most importantly, fragments will give rise to specialized inference techniques that can make automated reasoning more effective.

A Cross Entropy in Logarithmic Real-Closed Fields

In this appendix we prove that the most important properties of *CE* and *CE*-minimization in the reals carry over to the general case of *CE* in arbitrary lrc-fields. We partition these results into two groups: the first group describes qualitative properties that can be derived on the basis of the axioms LRCF without the approximation schema *(viii)*. The second group deals with the numerical agreement between *CE* in the reals and in other lrc-fields, and is essentially based on the schema LRCF*(viii)*.

A.1 Qualitative Properties

Lemma A.1 The following sentences are derivable from LRCF.

$$\mathit{Log}(1) = 0 \tag{92}$$

$$\forall x > 0 \;\; \mathit{Log}(1/x) = -\mathit{Log}(x) \tag{93}$$

$$\forall x \in (0,1) \;\; \mathit{Log}(x) < 0 \tag{94}$$

$$\forall x > 1 \;\; \mathit{Log}(x) > 0 \tag{95}$$

$$\forall x, y > 0 \;\; x < y \rightarrow \mathit{Log}(x) < \mathit{Log}(y) \tag{96}$$

$$0 \cdot \mathit{Log}(0) = 0 \tag{97}$$

The proofs for (92)-(96) are straightforward from the axioms LRCF. For (97) note that in every model $\mathfrak{F}$ for $\mathrm{S_{LOF}}$ a value $\mathit{Log}(0) \in \mathbb{F}$ has to be defined, and that by the field axioms $0 \cdot \mathit{Log}(0) = 0$ must hold. [4]

The following property of the logarithm is the basis for all the subsequent results in this section.

Lemma A.2 In every lrc-field the following holds:

$$\forall x_1, y_1, x_2, y_2 > 0 : \;\; x_1 \mathit{Log}\left(\frac{x_1}{y_1}\right) + x_2 \mathit{Log}\left(\frac{x_2}{y_2}\right) \geq (x_1 + x_2) \mathit{Log}\left(\frac{x_1 + x_2}{y_1 + y_2}\right), \tag{98}$$

where equality holds iff

$$\frac{x_1}{x_1 + x_2} = \frac{y_1}{y_1 + y_2}. \tag{99}$$

Proof: Let $\mathbb{F}$ be an lrc-field, and $x_1, y_1, x_2, y_2 \in \mathbb{F}$ be positive. Defining

$$\begin{array}{ll} x := x_1 + x_2, & \lambda_x := \frac{x_1}{x_1 + x_2} \\ y := y_1 + y_2, & \lambda_y := \frac{y_1}{y_1 + y_2}, \end{array}$$

we can write

$$x_1 = \lambda_x x, \quad x_2 = (1 - \lambda_x) x, \quad y_1 = \lambda_y y, \quad y_2 = (1 - \lambda_y) y,$$

[4] For $\mathbb{R}$ to be a formal model of LRCF one would have to define (arbitrary) values $\mathit{Log}(x) \in \mathbb{R}$ for $x \leq 0$. Note that in $\mathbb{R}$ the otherwise somewhat artificial identity (97) is given real meaning by the fact that $\mathit{lim}_{x \rightarrow 0} x \mathit{Log}(x) = 0$.

and the left hand side of (98) may be rewritten as

$$\begin{aligned} & \lambda_x x \mathit{Log}\left(\frac{\lambda_x x}{\lambda_y y}\right) + (1-\lambda_x) x \mathit{Log}\left(\frac{(1-\lambda_x)x}{(1-\lambda_y)y}\right) \\ = \; & x\mathit{Log}\left(\frac{x}{y}\right) + x\left(\lambda_x \mathit{Log}\left(\frac{\lambda_x}{\lambda_y}\right) + (1-\lambda_x)\mathit{Log}\left(\frac{1-\lambda_x}{1-\lambda_y}\right)\right). \end{aligned} \tag{100}$$

If (99) holds, i.e. $\lambda_x = \lambda_y$, then the second term of (100) vanishes by (92), so that (98) holds with equality.

Now suppose that $\lambda_x \neq \lambda_y$. Then $\frac{\lambda_y}{\lambda_x} \neq 1$ and $\frac{1-\lambda_y}{1-\lambda_x} \neq 1$. By LRCF*(v)*, $-\mathit{Log}(x) > 1 - x$ for $x \neq 1$, so that

$$\begin{aligned} \lambda_x \mathit{Log}\left(\frac{\lambda_x}{\lambda_y}\right) + (1-\lambda_x)\mathit{Log}\left(\frac{1-\lambda_x}{1-\lambda_y}\right) &= \\ \lambda_x\left(-\mathit{Log}\left(\frac{\lambda_y}{\lambda_x}\right)\right) + (1-\lambda_x)\left(-\mathit{Log}\left(\frac{1-\lambda_y}{1-\lambda_x}\right)\right) &> \\ \lambda_x\left(1-\frac{\lambda_y}{\lambda_x}\right) + (1-\lambda_x)\left(1-\frac{1-\lambda_y}{1-\lambda_x}\right) &= 0. \end{aligned}$$

Since $x > 0$, this means that the second term of (100) is strictly greater than 0. This proves the lemma.

□

Lemma A.3 (Positivity) Let $\mathfrak{F}$ be an lrc-field, $n \geq 2$, $Q, P \in \Delta^n_{\mathbb{F}}$ with $Q \ll P$. Then $\mathit{CE}(Q,P) \geq 0$, with equality iff $Q = P$.

Proof: By induction on n. Let $n = 2$, $Q = (Q_1, Q_2), P = (P_1, P_2) \in \Delta^2_{\mathbb{F}}$, $Q \ll P$. If one of the P_i equals 0, then so does the corresponding Q_i, in which case $Q = P$ and $\mathit{CE}(Q,P) = 1\mathit{Log}(1) = 0$. Suppose, then, that $P_i > 0$ $(i = 1, 2)$. If $Q_i = 0$ for one i, say $i = 1$, then $Q \neq P$ and $\mathit{CE}(Q,P) = \mathit{Log}\left(\frac{1}{P_2}\right) > 0$ by (95).

For the case that $Q_i, P_i > 0$ $(i = 1, 2)$, we have

$$\begin{aligned} \mathit{CE}(Q,P) &= Q_1 \mathit{Log}(\frac{Q_1}{P_1}) + Q_2 \mathit{Log}(\frac{Q_2}{P_2}) \\ &\geq (Q_1+Q_2)\mathit{Log}(\frac{Q_1+Q_2}{P_1+P_2}) \\ &= 1\mathit{Log}(1) \;=\; 0 \end{aligned}$$

by lemma A.2, with equality iff $Q_1/(Q_1+Q_2) = P_1/(P_1+P_2)$, i.e. $Q = P$.

Now let $n > 2$, and assume that the lemma has been shown for $n - 1$. For $Q = P$ we again obtain $\mathit{CE}(Q,P) = 1\mathit{Log}(1) = 0$. Suppose, then, that $Q \neq P$. Without loss of generality, $Q_1 \neq P_1$. Define $\bar{Q}, \bar{P} \in \Delta^{n-1}_{\mathbb{F}}$ by

$$\bar{Q}_i := Q_i \quad \bar{P}_i := P_i \quad i = 1, \ldots, n-2,$$

and

$$\bar{Q}_{n-1} := Q_{n-1} + Q_n \quad \bar{P}_{n-1} := P_{n-1} + P_n.$$

Then $\bar{Q} \ll \bar{P}$, $\bar{Q} \neq \bar{P}$, so that by induction hypothesis $\mathit{CE}(\bar{Q}, \bar{P}) > 0$. By lemma A.2 we have $\mathit{CE}(Q, P) \geq \mathit{CE}(\bar{Q}, \bar{P})$, which proves the lemma.

□

Lemma A.4 (Convexity) Let $\mathfrak{F}$ be an lrc-field, $n \geq 2$, $Q, Q', P \in \Delta^n_{\mathbb{F}}$, $Q \neq Q'$ with $Q, Q' \ll P$. Let $0 < \lambda < 1$. Then

$$\mathit{CE}(\lambda Q + (1-\lambda)Q', P) < \lambda \mathit{CE}(Q, P) + (1-\lambda)\mathit{CE}(Q', P).$$

Proof: For the proof of the lemma it is sufficient to show that for fixed $y \in \mathbb{F}$, $y > 0$, the function

$$c_y : \quad x \mapsto x \mathit{Log}\left(\frac{x}{y}\right)$$

defined for $x \geq 0$ is strictly convex, because then

$$\begin{aligned} \mathit{CE}(\lambda Q + (1-\lambda)Q', P) &= \sum_{P_i > 0} c_{P_i}(\lambda Q_i + (1-\lambda)Q'_i) \\ &< \sum_{P_i > 0} \lambda c_{P_i}(Q_i) + (1-\lambda)c_{P_i}(Q'_i) \\ &= \lambda \mathit{CE}(Q, P) + (1-\lambda)\mathit{CE}(Q', P), \end{aligned}$$

where the strict inequality holds because $Q_i \neq Q'_i$ for at least one $i \in \{1, \ldots, n\}$ with $P_i > 0$.

For the proof of the convexity of c_y, let $y > 0$, $x_1, x_2 \geq 0$, $x_1 \neq x_2$, $0 < \lambda < 1$. Abbreviate $\lambda x_1 + (1-\lambda)x_2$ by $\bar{x}$.

We distinguish two cases: first assume that one of the x_i is equal to 0, e.g. $x_1 = 0$. Then

$$\begin{aligned} c_y(\bar{x}) &= (1-\lambda)x_2 \mathit{Log}\left(\frac{(1-\lambda)x_2}{y}\right) \\ &< (1-\lambda)x_2 \mathit{Log}\left(\frac{x_2}{y}\right) \\ &= \lambda c_y(x_1) + (1-\lambda)c_y(x_2), \end{aligned}$$

where the inequality is due to (96), and the final equality holds because $c_y(0) = 0$ by (97).

Now suppose that $x_1, x_2 > 0$. By lemma A.2 we obtain

$$c_y(\bar{x}) \leq \lambda x_1 \mathit{Log}\left(\frac{\lambda x_1}{y/2}\right) + (1-\lambda)x_2 \mathit{Log}\left(\frac{(1-\lambda)x_2}{y/2}\right) \tag{101}$$

with equality iff $\lambda x_1/\bar{x} = 1/2$, i.e.

$$\lambda x_1 = (1-\lambda)x_2. \tag{102}$$

The right side of (101) may be rewritten as

$$\lambda x_1 \mathit{Log}\left(\frac{x_1}{y}\right) + \lambda x_1 \mathit{Log}(2\lambda) + (1-\lambda)x_2 \mathit{Log}\left(\frac{x_2}{y}\right) + (1-\lambda)x_2 \mathit{Log}(2(1-\lambda)).$$

Without loss of generality, assume that $\lambda x_1 \geq (1-\lambda)x_2$, so that we obtain

$$c_y(\bar{x}) \leq \lambda c_y(x_1) + (1-\lambda)c_y(x_2) + \lambda x_1 \mathit{Log}(4\lambda(1-\lambda)), \tag{103}$$

still with equality iff (102) holds.

First consider the case that (102) in fact is true. Then, because $x_1 \neq x_2$, we have that $\lambda \neq 1/2$. By the completeness of RCF, and the fact that

$$\mathbb{R} \models \forall \lambda \in (0,1) \quad \lambda \neq \frac{1}{2} \rightarrow \lambda \cdot (1-\lambda) < \frac{1}{4},$$

we infer that $4\lambda(1-\lambda) < 1$, which (with (94)) entails that $\lambda x_1 \mathit{Log}(4\lambda(1-\lambda)) < 0$, thus proving that

$$c_y(\bar{x}) < \lambda c_y(x_1) + (1-\lambda)c_y(x_2). \tag{104}$$

In almost the same manner (104) is derived for the case that (102) does not hold: the last term in (103) then is found to be ≤ 0, which suffices to prove (104) because we have strict inequality in (103).

□

So far we have established properties of *CE* as a function. Next we turn to the process of *CE*-minimization. The following two theorems state two key structural properties of cross-entropy minimization. These properties are the cornerstones of Shore's and Johnson's (1980) axiomatic justification of cross-entropy minimization, and, in a somewhat different guise, also of Paris's and Vencovská's (1990) derivation of the maximum entropy principle.

Theorem A.5 (System Independence) Let $\mathfrak{A}, \mathfrak{A}'$ be finite algebras. Let $\mathbb{F}$ be an lrc-field, $J \cup \{P\} \subseteq \Delta_{\mathbb{F}}\mathfrak{A}$, $J' \cup \{P'\} \subseteq \Delta_{\mathbb{F}}\mathfrak{A}'$. Define

$$\mathfrak{A}^{\times} := \mathfrak{A} \times \mathfrak{A}', \quad P^{\times} := P \otimes P',$$

and let $J^{\times} \subseteq \mathfrak{A}^{\times}$ be defined as the set of measures with marginal distribution on $\mathfrak{A}$ in J and marginal distribution on $\mathfrak{A}'$ in J', i.e.

$$J^{\times} = \{Q^{\times} \in \Delta_{\mathbb{F}}\mathfrak{A}^{\times} \mid \ Q^{\times} \upharpoonright \mathfrak{A} \in J, \ Q^{\times} \upharpoonright \mathfrak{A}' \in J'\}.$$

Then

$$\Pi_{J^{\times}}(P^{\times}) = \Pi_J(P) \otimes \Pi_{J'}(P') := \{Q \otimes Q' \mid Q \in \Pi_J(P), Q' \in \Pi_{J'}(P')\}. \tag{105}$$

Having established lemmas A.1-A.4, the proof of this theorem and the following can be carried out for lrc-field valued probabilities just as for real-valued probabilities. We will therefore omit the proofs here, and refer the reader to (Shore & Johnson 1980) and (Jaeger 1995*a*).

Theorem A.6 (Subset Independence) Let $\mathfrak{A}$ be a finite algebra on M, $A = \{A_1, \ldots, A_L\} \subseteq \mathfrak{A}$ a partition of M, and $\mathfrak{F}$ an lrc-field. Let $P \in \Delta_{\mathbb{F}}\mathfrak{A}$.

Denote by $\bar{\mathfrak{A}}$ the subalgebra of $\mathfrak{A}$ generated by A, and by $\mathfrak{A}^h$ the relative algebra of $\mathfrak{A}$ with respect to A_h $(h = 1, \ldots, L)$. For $Q \in \Delta_{\mathbb{F}}\mathfrak{A}$ let $\bar{Q}$ denote the restriction $Q \restriction \bar{\mathfrak{A}}$, and Q^h the conditional of Q on $\mathfrak{A}^h$ $(h = 1, \ldots, L;\ Q(A_h) > 0)$.

Let $J \subseteq \Delta_{\mathbb{F}}\mathfrak{A}$ be of the form

$$J = \bar{J} \cap J_1 \cap \ldots \cap J_L$$

with $\bar{J}$ a set of constraints on $\bar{Q}$, and J_h a set of constraints on Q^h $(h = 1, \ldots, L)$. Precisely:

$$\begin{array}{ll} \bar{J} = \{Q \in \Delta_{\mathbb{F}}\mathfrak{A} \mid \bar{Q} \in \bar{J}^*\} & \text{for some } \bar{J}^* \subseteq \Delta_{\mathbb{F}}\bar{\mathfrak{A}}, \\ J_h = \{Q \in \Delta_{\mathbb{F}}\mathfrak{A} \mid Q(A_h) = 0 \vee Q^h \in J_h^*\} & \text{for some } J_h^* \subseteq \Delta_{\mathbb{F}}\mathfrak{A}^h. \end{array}$$

Let $Q \in \Pi_J(P)$. For all $h \in \{1, \ldots, L\}$ with $Q(A_h) > 0$ then

$$Q^h \in \Pi_{J_h^*}(P^h). \tag{106}$$

An important consequence of theorem A.6 is that in the special case where J is defined by prescribing fixed probability values for the elements of a partition of M, then cross-entropy minimization reduces to *Jeffrey's rule* (Jeffrey 1965):

Corollary A.7 (Jeffrey's Rule) Let $\mathfrak{A}$ be a finite algebra on M, $P \in \Delta_{\mathbb{F}}\mathfrak{A}$, $\{A_1, \ldots, A_L\} \subset \mathfrak{A}$ a partition of M, and $(r_1, \ldots, r_L) \in \Delta^L_{\mathbb{F}}$ with $r_h > 0 \Rightarrow P(A_h) > 0$ for $h = 1, \ldots, L$. For

$$J := \{Q \in \Delta_{\mathbb{F}}\mathfrak{A} \mid Q(A_h) = r_h;\ h = 1, \ldots, L\}$$

then $\Pi_J(P) = \{Q\}$ for

$$Q = \sum_{\substack{h = 1 \\ r_h > 0}}^{L} r_h P^h, \tag{107}$$

where P^h is the conditional of P on A_h.

A.2 Numerical Approximations

To motivate the results in this section, reconsider the example of section 3.1 given by (46)-(50). Here (47)-(50) defined a unique statistical probability measure $P = (0.4, 0.3, 0.1, 0.2)$ on a four-element algebra. The components of P being rational,

P can be interpreted as an element $P(\mathbb{F})$ of $\Delta^4_{\mathbb{F}}$ for any lrc-field $\mathbb{F}$. Similarly, the constraint (46) defines a subset

$$J(\mathbb{F}) := \{(x_1, \ldots, x_4) \in \Delta^4_{\mathbb{F}} \mid x_1 + x_2 = 0.5,\ x_2 + x_3 = 0.7\}$$

of $\Delta^4_{\mathbb{F}}$ for every $\mathbb{F}$. For the inductive inference relation of $\mathscr{L}_{ip}$ we now have to consider the *CE*-projections $\Pi_{J(\mathbb{F})}(P(\mathbb{F}))$ for arbitrary $\mathbb{F}$. For $\mathbb{F} = \mathbb{R}$ we know that $\Pi_{J(\mathbb{F})}(P(\mathbb{F}))$ contains a unique element Q, and, using an iterative nonlinear optimization algorithm, we can determine the value of Q approximately, as stated in (51). More precisely, the meaning of (51) is

$$\Pi_{J(\mathbb{R})}(P(\mathbb{R})) \subseteq \{(q_1, \ldots, q_4) \in \Delta^4_{\mathbb{R}} \mid q_1 \in (0.128, 0.129), \ldots, q_4 \in (0.171, 0.172)\}. \tag{108}$$

In order to use this numerical result obtained for the reals for showing that certain inductive entailment relations hold in $\mathscr{L}_{ip}$ – e.g. that (52) follows from (46)-(50) – we have to ascertain that (108) implies

$$\Pi_{J(\mathbb{F})}(P(\mathbb{F})) \subseteq \{(q_1, \ldots, q_4) \in \Delta^4_{\mathbb{F}} \mid q_1 \in (0.128, 0.129), \ldots, q_4 \in (0.171, 0.172)\} \tag{109}$$

for every $\mathbb{F}$. Theorem A.10 will show that this is indeed the case. We obtain this result by showing successively that the bounds given for *Log* by LRCF*(viii)* are sufficient to determine uniform bounds (i.e. valid in every $\mathbb{F}$) for the function $x\mathit{Log}(x/q)$ ($q \in \mathbb{Q}$ fixed), for $\mathit{CE}(Q, P)$ ($P \in \Delta^n_{\mathbb{Q}}$ fixed), and finally for $\Pi_{J(\mathbb{F})}(P(\mathbb{F}))$. The first lemma gives a piecewise approximation of $x\mathit{Log}(x/q)$.

Lemma A.8 Let $\epsilon > 0$ and $P \in (0, 1]$ be rational numbers [5], let p_n and q_n be as defined in LRCF(viii). There exists a rational number $r(\epsilon) > 0$ and an $m \in \mathbb{N}$ such that the following $\mathrm{S}_{\mathrm{LOF}}$-sentences hold in all lrc-fields:

$$\forall x \in (0, r(\epsilon)]\ x\mathit{Log}(\frac{x}{P}) \in (-\epsilon, 0) \tag{110}$$

$$\forall x \in [r(\epsilon), P]\ x\mathit{Log}(\frac{x}{P}) \in [xq_m(\frac{x}{P}), xp_m(\frac{x}{P})] \tag{111}$$

$$\forall x \in [r(\epsilon), P]\ xp_m(\frac{x}{P}) - xq_m(\frac{x}{P}) \in [0, \epsilon) \tag{112}$$

$$\forall x \in [P, 1]\ x\mathit{Log}(\frac{x}{P}) \in [-xp_m(\frac{P}{x}), -xq_m(\frac{P}{x})] \tag{113}$$

$$\forall x \in [P, 1]\ - xq_m(\frac{P}{x}) + xp_m(\frac{P}{x}) \in [0, \epsilon). \tag{114}$$

Proof: We first determine a number $r(\epsilon)$ such that the approximation (110) holds. We then choose a sufficiently large n such that the bounds (112) and (114) hold. Properties (111) and (113) directly follow from LRCF(viii).

By elementary calculus we find that in $\mathbb{R}$ $\mathit{lim}_{x \to 0} x\mathit{Log}(\frac{x}{P}) = 0$, and that $x\mathit{Log}(\frac{x}{P})$ attains its absolute minimum at $x = \frac{P}{e} > 0$.

[5] All the results in this section remain valid when we substitute “algebraic numbers” for “rational numbers” throughout.

We choose an arbitrary rational $r(\epsilon) \in (0, \frac{P}{e})$ that satisfies

$$r(\epsilon)Log(\frac{r(\epsilon)}{P}) > max\{-\epsilon, \frac{P}{e}Log(\frac{P}{e})\}$$

Also, choose a rational $r' \in (r(\epsilon), \frac{P}{e})$. By the strict convexity of $x \mapsto xLog(x/P)$ then $r'Log(\frac{r'}{P}) < r(\epsilon)Log(\frac{r(\epsilon)}{P})$. For sufficiently large $n \in \mathbb{N}$

$$r(\epsilon)q_m(\frac{r(\epsilon)}{P}) > r'p_m(\frac{r'}{P}) \quad \text{and} \quad r(\epsilon)q_m(\frac{r(\epsilon)}{P}) > -\epsilon$$

now holds in $\mathbb{R}$, and hence in every lrc-field. It follows that in every lrc-field we have

$$r(\epsilon)Log(\frac{r(\epsilon)}{P}) > r'Log(\frac{r'}{P}) \quad \text{and} \quad r(\epsilon)Log(\frac{r(\epsilon)}{P}) > -\epsilon.$$

By the strict convexity of the function $x \mapsto xLog(\frac{x}{P})$ (lemma A.4) we can now infer

$$\forall x \in (0, r(\epsilon)]\ xLog(\frac{x}{P}) > r(\epsilon)Log(\frac{r(\epsilon)}{P}),$$

and thus

$$\forall x \in (0, r(\epsilon)]\ xLog(\frac{x}{P}) > -\epsilon.$$

Also, because $r(\epsilon) < P$, by (93) and (95) we get

$$\forall x \in (0, r(\epsilon)]\ xLog(\frac{x}{P}) < 0,$$

proving (110).

For the approximation of $xLog(\frac{x}{P})$ on $[r(\epsilon), 1]$ choose an $m \in \mathbb{N}$ such that

$$max\{\frac{(r(\epsilon)-1)^{m+1}}{r(\epsilon)}, \frac{(P-1)^{m+1}}{P}\} < \epsilon.$$

For such m then (112) and (114) are satisfied.

□

The next lemma combines bounds for $Q_iLog(Q_i/P_i)$ to find bounds for $CE(Q,P)$. In the formulation of the lemma we employ the notations introduced in section 2.3 for the interpretations of terms in a structure, and for the sets defined in a structure by a formula.

Lemma A.9 Let $n \geq 1$, $P \in \Delta^n_{\mathbb{Q}}$, and $\epsilon \in \mathbb{Q}$, $\epsilon > 0$. There exist $L_I(\mathrm{S_{OF}})$-formulas $\alpha_1(\boldsymbol{x}), \ldots, \alpha_k(\boldsymbol{x})$ and $L_I(\mathrm{S_{OF}})$-terms $l_1(\boldsymbol{x}), u_1(\boldsymbol{x}), \ldots, l_k(\boldsymbol{x}), u_k(\boldsymbol{x})$ with $\boldsymbol{x} = (x_1, \ldots, x_n)$, such that the following holds in all lrc-fields $\mathfrak{F}$:

(i) $\Delta^n_{\mathbb{F}} \cap \{Q \mid Q \ll P\} = \cup_{i=1}^k (\mathfrak{F}, \boldsymbol{x})(\alpha_i)$

(ii) $\forall i \in \{1, \ldots, k\} \forall Q \in (\mathfrak{F}, \boldsymbol{x})(\alpha_i) :\ \mathfrak{F}(l_i(Q)) \leq CE(Q,P) \leq \mathfrak{F}(u_i(Q))$, and $\mathfrak{F}(u_i(Q)) - \mathfrak{F}(l_i(Q)) < \epsilon$.

Proof: Let $P \in \Delta^n_{\mathbb{Q}}$. Assume, first, that $P_i > 0$ for all $i = 1, \ldots, n$, so that $Q \ll P$ for all $Q \in \Delta^n_{\mathbb{F}}$. Applying lemma A.8 to the P_i and ϵ/n, we find rational constants $r_1(\epsilon/n), \ldots, r_n(\epsilon/n)$, such that $Q_i Log \frac{Q_i}{P_i}$ can be bounded for $Q_i \in (0, r_i(\epsilon/n)]$ by the constants $-\epsilon/n$ and 0, and for $Q_i \in [r_i(\epsilon/n), 1]$ by the terms $Q_i q_m(Q_i/P_i), Q_i q_m(P_i/Q_i), Q_i p_m(Q_i/P_i), Q_i p_m(P_i/Q_i)$ as described in lemma A.8.

We now let the formulas α_j run over all conjunctions of the form

$$\wedge_{i=1}^n (x_i \in I_i),$$

where I_i is either $(0, r_i(\epsilon/n)]$, $[r_i(\epsilon/n), P_i]$, or $[P_i, 1]$. The lower bound $l_j(\boldsymbol{x})$ on $CE(Q, P)$ for elements Q of $\alpha_j(\boldsymbol{x})$ then is given by the sum of the lower bounds $-\epsilon/n$, $Q_i q_m(Q_i/P_i)$, respectively $-Q_i p_m(P_i/Q_i)$, obtained for each component $Q_i Log \frac{Q_i}{P_i}$ of $CE(Q, P)$. Similarly for the upper bounds $u_j(\boldsymbol{x})$.

If $P_i = 0$ for some $i \in \{1, \ldots, n\}$ we proceed in the same way, simply using a conjunct $x_i = 0$ instead of a conjunct $x_i \in I_i$ in the definition of the α_j.

□

Now the desired theorem can be formulated. Roughly speaking, it says that approximations of the *CE*-projection $\Pi_J(P)$ that are expressible by a $\mathrm{S_{OF}}$-formula, and that are valid in $\mathbb{R}$, also are valid in arbitrary $\mathbb{F}$.

Theorem A.10 Let $\phi(x_1, \ldots, x_n)$ and $\psi(x_1, \ldots, x_n)$ be $L_I(\mathrm{S_{OF}})$-formulas. Let $P \in \Delta^n_{\mathbb{Q}}$. Define

$$\chi(\phi, \psi) :\equiv \exists x > 0 \exists \boldsymbol{z}(\phi(\boldsymbol{z}) \wedge \forall \boldsymbol{y}(\phi(\boldsymbol{y}) \wedge \neg\psi(\boldsymbol{y}) \rightarrow CE(\boldsymbol{z}, P) < CE(\boldsymbol{y}, P) - x)).$$

If $\mathfrak{R} \models \chi(\phi, \psi)$, then LRCF $\models \chi(\phi, \psi)$.

To connect this theorem with our introductory example, think of ϕ as the formula defining the set $J(\mathbb{F})$ and of ψ as the formula defining the right-hand side of (109). Then $\chi(\phi, \psi)$ essentially is the general statement whose interpretation over $\mathbb{R}$ is (108), and whose interpretation over $\mathbb{F}$ is (109). The theorem now says that (108) implies (109).

Proof: Assume that $\mathbb{R} \models \chi(\phi, \psi)$, and let $0 < \epsilon \in \mathbb{Q}$ be such that $\mathbb{R}$ is a model of

$$\exists \boldsymbol{z}(\phi(\boldsymbol{z}) \wedge \forall \boldsymbol{y}(\phi(\boldsymbol{y}) \wedge \neg\psi(\boldsymbol{y}) \rightarrow CE(\boldsymbol{z}, P) < CE(\boldsymbol{y}, P) - \epsilon)). \tag{115}$$

Let $\alpha_1(\boldsymbol{x}), \ldots, \alpha_k(\boldsymbol{x})$ and $l_1(\boldsymbol{x}), u_1(\boldsymbol{x}), \ldots, l_k(\boldsymbol{x}), u_k(\boldsymbol{x})$ be as given by lemma A.9 for P and $\epsilon/3$. Then, for some $j \in \{1, \ldots, k\}$, $\mathbb{R}$ also is a model of

$$\begin{aligned} \exists \boldsymbol{z}(\phi(\boldsymbol{z}) \wedge \alpha_j(\boldsymbol{z}) \wedge & \forall \boldsymbol{y} \ll P \exists i \in \{1, \ldots, k\} \\ & (\alpha_i(\boldsymbol{y}) \wedge (\phi(\boldsymbol{y}) \wedge \neg\psi(\boldsymbol{y}) \rightarrow u_j(\boldsymbol{z}) < l_i(\boldsymbol{y}) - \epsilon/3))), \end{aligned} \tag{116}$$

which, some abuse of first-order syntax notwithstanding, is a pure $L_I(\mathrm{S_{OF}})$-sentence. Thus, (116) holds in every lrc-field $\mathfrak{F}$.

Furthermore, by lemma A.9, we have for arbitrary $\mathfrak{F}$:

$$\mathfrak{F} \models \forall \boldsymbol{y} \forall i \in \{1, \ldots, k\}(\alpha_i(\boldsymbol{y}) \rightarrow \mathit{CE}(\boldsymbol{y}, P) - l_i(\boldsymbol{y}) \in [0, \epsilon/3] \wedge \\ u_i(\boldsymbol{y}) - \mathit{CE}(\boldsymbol{y}, P) \in [0, \epsilon/3]). \qquad (117)$$

Combining the bounds $l_i(\boldsymbol{y}) - u_j(\boldsymbol{z}) > \epsilon/3$, $\mathit{CE}(\boldsymbol{y}, P) - l_i(\boldsymbol{y}) \leq \epsilon/3$, and $u_j(\boldsymbol{z}) - \mathit{CE}(\boldsymbol{z}, P) \leq \epsilon/3$, one obtains $\mathit{CE}(\boldsymbol{y}, P) - \mathit{CE}(\boldsymbol{z}, P) > \epsilon/3$, so that (115) with ϵ replaced by $\epsilon/3$ holds in arbitrary $\mathfrak{F}$, and hence also $\mathfrak{F} \models \chi(\phi, \psi)$.

□

The following corollary mediates between the rather abstract formulation of theorem A.10 and our introductory example.

Corollary A.11 Let $J \subseteq \Delta^n_{\mathbb{R}}$ be closed and defined by an $L_I(\mathrm{S}_{\mathrm{OF}})$-formula $\phi(x_1, \ldots, x_n)$. Let $H \subseteq \Delta^n_{\mathbb{R}}$ be open and defined by an $L_I(\mathrm{S}_{\mathrm{OF}})$-formula $\psi(x_1, \ldots, x_n)$. Let $P \in \Delta^n_{\mathbb{Q}}$, and assume that $\Pi_J(P) \subset H$. For an arbitrary lrc-field $\mathfrak{F}$, and the sets $\bar{J}, \bar{H}$ defined in $\mathfrak{F}$ by ϕ and ψ, respectively, then $\Pi_{\bar{J}}(P) \subset \bar{H}$.

Proof: According to the assumptions the set $H^c \cap J$ is closed. Let $Q \in \Pi_J(P)$. From $\Pi_J(P) \subset H$ and the compactness of $H^c \cap J$ it follows that there exists $\epsilon \in \mathbb{R}^+$ such that $\mathit{CE}(Q, P) < \mathit{CE}(Q', P) - \epsilon$ for every $Q' \in H^c \cap J$. Thus $\mathfrak{R} \models \chi(\phi, \psi)$. By theorem A.10 then $\mathfrak{F} \models \chi(\phi, \psi)$, which entails $\Pi_{\bar{J}}(P) \subset \bar{H}$.

□

Acknowledgments

The author thanks the anonymous reviewers of an earlier version of this paper for their perceptive and constructive comments. Specifically, comments and examples provided by one reviewer have been integrated into sections 2.4 and 3.3 of the present paper.

References

Abadi, M. & J.Y.Halpern (1994), 'Decidability and expressiveness for first-order logics of probability', *Information and Computation* **112**, 1–36.

Bacchus, F. (1990*a*), 'Lp, a logic for representing and reasoning with statistical knowledge', *Computational Intelligence* **6**, 209–231.

Bacchus, F. (1990*b*), *Representing and Reasoning With Probabilistic Knowledge*, MIT Press.

Bacchus, F., Grove, A., Halpern, J. & Koller, D. (1992), From statistics to beliefs, *in* 'Proc. of National Conference on Artificial Intelligence (AAAI-92)'.

Bacchus, F., Grove, A. J., Halpern, J. Y. & Koller, D. (1997), 'From statistical knowledge bases to degrees of belief', *Artificial Intelligence* **87**, 75–143.

Boole, G. (1854), *Investigations of Laws of Thought on which are Founded the Mathematical Theories of Logic and Probabilities*, London.

Carnap, R. (1950), *Logical Foundations of Probability*, The University of Chicago Press.

Carnap, R. (1952), *The Continuum of Inductive Methods*, The University of Chicago Press.

Dahn, B. I. & Wolter, H. (1983), 'On the theory of exponential fields', *Zeitschrift für mathematische Logik und Grundlagen der Mathematik* **29**, 465–480.

de Finetti, B. (1937), 'La prévision: ses lois logiques, ses sources subjectives', *Annales de l'Institut Henri Poincaré*. English Translation in (Kyburg & Smokler 1964).

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

Diaconis, P. & Zabell, S. (1982), 'Updating subjective probability', *Journal of the American Statistical Association* **77**(380), 822–830.

Dubois, D. & Prade, H. (1997), Focusing vs. belief revision: A fundamental distinction when dealing with generic knowledge, *in* 'Proceedings of the First International Joint Conference on Qualitative and Quantitative Practical Reasoning', Springer-Verlag, pp. 96–107.

Fenstad, J. E. (1967), Representations of probabilities defined on first order languages, *in* J. N. Crossley, ed., 'Sets, Models and Recursion Theory', North Holland, Amsterdam, pp. 156–172.

Gaifman, H. (1964), 'Concerning measures in first order calculi', *Israel Journal of Mathematics*.

Gaifman, H. & Snir, M. (1982), 'Probabilities over rich languages, testing and randomness', *Journal of Symbolic Logic* **47**(3), 495–548.

Gilboa, I. & Schmeidler, D. (1993), 'Updatin ambiguous beliefs', *Journal of Economic Theory* **59**, 33–49.

Grove, A. & Halpern, J. (1998), Updating sets of probabilities, *in* 'Proceedings of the Fourteenth Conference on Uncertainty in AI', pp. 173–182.

Grove, A., Halpern, J. & Koller, D. (1992*a*), Asymptotic conditional probabilities for first-order logic, *in* 'Proc. 24th ACM Symp. on Theory of Computing'.

Grove, A., Halpern, J. & Koller, D. (1992*b*), Random worlds and maximum entropy, *in* 'Proc. 7th IEEE Symp. on Logic in Computer Science'.

Hailperin, T. (1976), *Boole's Logic and Probability*, Vol. 85 of *Studies in Logic and the Foundations of Mathematics*, North-Holland.

Hailperin, T. (1996), *Sentential Probability Logic*, Lehigh University Press, Bethlehem.

Halpern, J. (1990), 'An analysis of first-order logics of probability', *Artificial Intelligence* **46**, 311–350.

Hoover, D. N. (1978), 'Probability logic', *Annals of Mathematical Logic* **14**, 287–313.

Jaeger, M. (1994*a*), A logic for default reasoning about probabilities, *in* R. Lopez de Mantaraz & D. Poole, eds, 'Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence (UAI'94)', Morgan Kaufmann, Seattle, USA, pp. 352–359.

Jaeger, M. (1994*b*), Probabilistic reasoning in terminological logics, *in* J. Doyle, E. Sandewall & P. Torasso, eds, 'Principles of Knowledge Representation an Reasoning: Proceedings of the 4th International Conference (KR94)', Morgan Kaufmann, Bonn, Germany, pp. 305–316.

Jaeger, M. (1995*a*), Default Reasoning about Probabilities, PhD thesis, Universität des Saarlandes.

Jaeger, M. (1995*b*), Minimum cross-entropy reasoning: A statistical justification, *in* C. S. Mellish, ed., 'Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)', Morgan Kaufmann, Montréal, Canada, pp. 1847–1852.

Jaynes, E. (1978), Where do we stand on maximum entropy?, *in* R. Levine & M. Tribus, eds, 'The Maximum Entropy Formalism', MIT Press, pp. 15–118.

Jeffrey, R. (1965), *The Logic of Decision*, McGraw-Hill.

Jensen, F. (2001), *Bayesian Networks and Decision Graphs*, Springer.

Keisler, H. (1985), Probability quantifiers, *in* J. Barwise & S. Feferman, eds, 'Model-Theoretic Logics', Springer-Verlag, pp. 509–556.

Kullback, S. (1959), *Information Theory and Statistics*, Wiley.

Kullback, S. & Leibler, R. A. (1951), 'On information and sufficiency', *Annals of mathematical statistics* **22**, 79–86.

Kyburg, H. E. (1974), *The Logical Foundations of Statistical Inference*, D. Reidel Publishing Company.

Kyburg, H. E. (1983), 'The reference class', *Philosophy of Science* **50**, 374–397.

Kyburg, H. E. & Smokler, H. E., eds (1964), *Studies in Subjective Probability*, John Wiley.

Lewis, D. (1976), 'Probabilities of conditionals and conditional probabilities', *The Philosophical Review* **85**(3), 297–315.

McCarthy, J. (1980), 'Circumscription - a form of non-monotonic reasoning', *Artificial Intelligence* **13**, 27–39.

Moral, S. & Wilson, N. (1995), Revision rules for convex sets of probabilities, *in* G. Coletti, D. Dubois & R. Scozzafava, eds, 'Mathematical Models for Handling Partial Knowledge in Artificial Intelligence', Kluwer.

Nilsson, N. (1986), 'Probabilistic logic', *Artificial Intelligence* **28**, 71–88.

Paris, J. & Vencovská, A. (1990), 'A note on the inevitability of maximum entropy', *International Journal of Approximate Reasoning* **4**, 183–223.

Paris, J. & Vencovská, A. (1992), 'A method for updating that justifies minimum cross entropy', *International Journal of Approximate Reasoning* **7**, 1–18.

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

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*, The Morgan Kaufmann series in representation and reasoning, rev. 2nd pr. edn, Morgan Kaufmann, San Mateo, CA.

Pollock, J. L. (1983), 'A theory of direct inference', *Theory and Decision* **15**, 29–95.

Rabin, M. O. (1977), Decidable theories, *in* J. Barwise, ed., 'Handbook of mathematical logic', Elsevier Science Publishers.

Reichenbach, H. (1949), *The Theory of Probability*, University of California Press.

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

Scott, D. & Krauss, P. (1966), Assigning probabilities to logical formulas, *in* J. Hintikka & P. Suppes, eds, 'Aspects of Inductive Logic', North Holland, Amsterdam, pp. 219–264.

Shafer, G. (1976), *A Mathematical Theory of Evidence*, Princeton University Press.

Shoham, Y. (1987), Nonmonotonic logics: Meaning and utility, *in* 'Proceedings of IJCAI-87'.

Shore, J. & Johnson, R. (1980), 'Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy', *IEEE Transactions on Information Theory* **IT-26**(1), 26–37.

Shore, J. & Johnson, R. (1983), 'Comments on and correction to "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy"', *IEEE Transactions on Information Theory* **IT-29**(6), 942–943.

von Mises, R. (1951), *Wahrscheinlichkeit Statistik und Wahrheit*, Springer.

von Mises, R. (1957), *Probability, Statistics and Truth*, George Allen & Unwin.

Walley, P. (1991), *Statistical Reasoning with Imprecise Probabilities*, Chapman and Hall.