Default Reasoning about Probabilities

Dissertation

zur Erlangung des Grades
des Doktors der Naturwissenschaften
der Technischen Fakultät
der Universität des Saarlandes
von

Manfred Jaeger

Saarbrücken, 1995

Acknowledgments

First and foremost I would like to thank Harald Ganzinger and Hans Jürgen Ohlbach who provided me with the opportunity to conduct this research at the Max-Planck-Institute for Computer Science. Here I found a perfect environment that gave me all the freedom I wanted and every support I needed.

Hans Jürgen put me on the track of probabilistic reasoning in terminological logics. The longer I followed this track, the more fascinating I found the questions I encountered, and the more convinced I became of doing just the research that suited me best. Alas, in this process I also left behind terminological logics and what Hans Jürgen originally had in mind that my thesis should be about.

More than anybody else, Emil Weydert has kept a close look on the progress of my work. While not sparing his critical comments, he showed an interest in this work that gave me much encouragement, and asked the questions that helped me find the right directions for my investigations.

I have much benefitted from the expertise of Martin Beibel and Daniel Hug in the fields of probability theory and convex geometry, as well as their direct access to the well-stocked mathematical library at the University of Freiburg.

At one stage of my work I shamelessly exploited Peter Barth's experience in C^{++}-programming. I don't know how I should ever have minimized a single cross-entropy without his help.

My roommate Luca Viganò has provided me with additional intellectual challenges by trying to teach me Italian and introducing me to a silly computer game called xjewel. In neither discipline have I come close to his mastery.

Contents

Zusammenfassung

Will man ein wissensbasiertes System entwickeln, das mit probabilistischer Information umgehen kann, wird man bald mit der Tatsache konfrontiert, daß es zwei grundlegend unterschiedliche Formen von Wahrscheinlichkeiten gibt, die beide ihren Platz in einem ausdrucksstarken System haben müssen: Zunächst gibt es die Erscheinungsform von Wahrscheinlichkeit als eine statistische Größe, die die relative Häufigkeit eines gewissen Merkmales in einer bestimmten Klasse von Objekten quantifiziert. Diesem statistischen Wahrscheinlichkeitsbegriff gegenüber steht der Begriff der subjektiven Wahrscheinlichkeit. Dieser wesentlich schwieriger zu präzisierende Begriff bezeichnet einen Grad der Überzeugung von der Wahrheit bestimmter Aussagen.

Ein Beispiel für eine statistische Wahrscheinlichkeit ist die Aussage "80% aller amerikanischen Kriminalfilme haben ein Happy End". Eine subjektive Wahrscheinlichkeit ist in der Aussage enthalten "Mit einer Wahrscheinlichkeit von mindestens 0.7 ist dieser Film der gerade im Fernsehen läuft eine amerikanische Produktion".

In der vorliegenden Arbeit wird ein ausdrucksstarker logischer Formalismus basierend auf der Prädikatenlogik erster Stufe definiert, welcher es erlaubt, Aussagen über beide Typen von Wahrscheinlichkeiten zu repräsentieren. Zentrales Anliegen bei der Definition der Semantik für die verwendete formale Sprache ist es, in ihr nicht nur das Ziehen wahrscheinlichkeitstheoretisch korrekter Schlüsse aus einer gegebenen Wissensbasis zu formalisieren, sondern auch plausible Inferenzen zu modellieren, mit denen ein Mensch subjektive Wahrscheinlichkeiten auf Grund von statistischer Information zuweist. In dem oben angegebenen Beispiel würde man etwa in aller Regel schließen, daß die Wahrscheinlichkeit für den gerade zu sehenden Film ein Happy End zu haben, mindestens $0.7 \cdot 0.8 = 0.56$ beträgt – obwohl kein Gesetz der Wahrscheinlichkeitstheorie diesen Schluß als korrekt legitimiert. Diesen Inferenzmechanismus haben wir hier "default reasoning about probabilities" genannt.

Kapitel 1 stellt zunächst einige grundlegende Begriffe bereit, die bei der Definition der Semantik für die zu entwickelnde Logik eine tragende Rolle spielen werden. Bemerkenswert ist hier vor allem, daß von dem in der Wahrscheinlichkeitstheorie üblichen Begriff eines Wahrscheinlichkeitsmaßes etwas abgegangen wird: Ein Wahrscheinlichkeitsmaß ist für uns eine Funktion die Werte in einem beliebigen reell abgeschlossenen Körper annehmen kann, nicht nur in den reellen Zahlen. Hierdurch werden wir später ein Vollständigkeitsresultat erzielen. Als notwendige Folge dieser Verallgemeinerung können Wahrscheinlichkeitsmaße dann nur noch das Axiom der endlichen Additivität erfüllen, nicht das der $\sigma-$Additivität.

Bevor wir uns der Handhabung von subjektiven Wahrscheinlichkeiten zuwenden, ist Kapitel 2 zunächst ganz den statistischen Wahrscheinlichkeiten gewidmet. Es baut weitgehend auf einer Logik zur Repräsentation von statistischer Wahrscheinlichkeit auf, die von Bacchus [1990a] entwickelt wurde. Insbesondere übernehmen wir hier unverändert die Syntax von Bacchus. Die Repräsentationssprache, hier L^σ genannt, entsteht aus der Sprache der Prädikatenlogik erster Stufe durch Hinzunahme statistischer Quantifizierung, die es erlaubt, aus einer Formel $\phi(\boldsymbol{v},\boldsymbol{w},\boldsymbol{x})$ mit Tupeln $\boldsymbol{v}$, $\boldsymbol{w}$, $\boldsymbol{x}$ von freien Variablen (wobei $\boldsymbol{v}$ und $\boldsymbol{w}$ für Variable stehen, die als Elemente aus der zu beschreibenden Menge von Objekten interpretiert werden, während $\boldsymbol{x}$ über Elemente eines reell abgeschlossenen Körpers, also über Wahrscheinlichkeiten, variiert) einen neuen Term

$$[\phi(\boldsymbol{v},\boldsymbol{w},\boldsymbol{x})]_{\boldsymbol{w}}$$

zu bilden.

Ein solcher Wahrscheinlichkeitsterm repräsentiert die statistische Wahrscheinlichkeit für ein zufällig gewähltes Tupel $\boldsymbol{a}$ von Objekten die Eigenschaft $\phi(\boldsymbol{v},\boldsymbol{a},\boldsymbol{x})$ aufzuweisen. Eine semantische Struktur, die es erlaubt diese anschauliche Bedeutung zu formalisieren, hat die Form

$$\mathfrak{M} = (M, I, \mathfrak{F}, (\mathfrak{A}_n, \mu_n)_n),$$

wobei M der Träger und I die Interpretationsfunktion einer gewöhnlichen modelltheoretischen Struktur für die zugrunde liegende Symbolmenge und $\mathfrak{F}$ ein reell abgeschlossener Körper ist. Für jedes $n \geq 1$ ist $\mathfrak{A}_n$ eine Algebra auf M^n und μ_n ein Wahrscheinlichkeitsmaß mit Werten in $\mathfrak{F}$ auf $\mathfrak{A}_n$. Um als Interpretation für die Sprache L^σ zu dienen, muß eine Struktur dieser allgemeinen Form noch eine Reihe von Bedingungen erfüllen. Zunächst werden für die Folge $(\mathfrak{A}_n, \mu_n)_n$ drei Konsistenzbedingungen verlangt, welche die wesentlichen Eigenschaften einer Folge von Produktmaßen verkörpern und somit sicherstellen, daß jedes Maß μ_n als die statistische Verteilung von n unabhängigen Stichproben aus dem Träger verstanden werden kann. Weiterhin müssen die Algebren $\mathfrak{A}_n$ eine Abschlußbedingung erfüllen, die sicherstellt, daß jede in der Sprache definierbare Teilmenge auch meßbar ist. Eine Struktur, die diese Bedingungen erfüllt, nennen wir eine statistische Struktur. In ihr wird ein Wahrscheinlichkeitsterm dann (innerhalb einer gewöhnlichen induktiven Definition der semantischen Modellbeziehung $\models$) durch das Element

$$\mu_{|\boldsymbol{w}|}(\{\boldsymbol{a} \mid \mathfrak{M} \models \phi(\boldsymbol{v},\boldsymbol{w},\boldsymbol{x})[\boldsymbol{w}/\boldsymbol{a}]\}) \in \mathfrak{F}$$

interpretiert. Die Logik, die aus der Sprache L^σ und der Modellbeziehung zwischen statistischen Strukturen und L^σ-Formeln besteht, nennen wir $\mathscr{L}^\sigma$.

Aufgrund des komplexen Aufbaus einer statistischen Struktur, insbesondere wegen der durch die Abschlußbedingung gegebenen Wechselwirkung zwischen der notwendigen algebraischen Struktur von $\mathfrak{A}_n$ und der Ausdrucksstärke der zu interpretierenden Sprache, ist es relativ schwierig, konkrete Beispiele für statistische Strukturen (abgesehen von einigen besonders einfachen Formen) anzugeben. In Abschnitt 2.4 wird deshalb eine allgemeine Konstruktion durchgeführt, die für den Spezialfall von monadischen Symbolmengen eine breite Klasse von statistischen Strukturen erschließt.

In Abschnitt 2.5 wird gezeigt, daß wir mit $\mathscr{L}^\sigma$ die Logik erster Stufe nicht wirklich verlassen haben. Es wird gezeigt, wie sich die Sprache L^σ in eine reine erststufige Sprache übersetzen läßt, indem statistische Quantoren in einem der Skolemisierung ähnlichen Verfahren durch neue Funktionssymbole ersetzt werden. Weiterhin läßt sich die Semantik der statistischen Quantoren in einem Axiomensystem für die neuen Funktionssymbole nachbilden. Dies wird durch die Tatsache ermöglicht, daß Wahrscheinlichkeiten in beliebigen reell abgeschlossenen Körpern angenommen werden können, also in einer in der Logik erster Stufe axiomatisierbaren Klasse von Strukturen. Die logische Folgerungsbeziehung bezüglich $\mathscr{L}^\sigma$ kann dann zurückgeführt werden auf die logische Folgerungsbeziehung in der Logik erster Stufe. Insbesondere überträgt sich die Vollständigkeit von Inferenzkalkülen für die Logik erster Stufe auf Inferenzen in $\mathscr{L}^\sigma$.

In Kapitel 3 verlassen wir zunächst den Rahmen formaler Logik und wenden uns der Analyse der Begriffe der subjektiven Wahrscheinlichkeit und des "default reasoning about probabilities" zu. Eine inhaltliche Klärung dieser Begriffe ist Voraussetzung für ihre Einbindung in ein formales logisches System.

Die Bedeutung einer subjektiven Wahrscheinlichkeit ist wesentlich schwieriger zu präzisieren als die einer statistischen Wahrscheinlichkeit. Klassische Interpretationen verwendeten meistens den Begriff einer fairen Gewinnquote bei einer Wette auf den Wahrheitsgehalt der in Frage stehenden Aussage. Wir schlagen hier einen anderen Weg ein. Es wird vorgeschlagen, die Angabe einer subjektiven Wahrscheinlichkeit aufzufassen als eine Voraussage über den Ausgang eines Gedankenexperimentes: Ordnen wir der Aussage, daß ein bestimmtes Ereignis e eine Eigenschaft ϕ besitzt, die Wahrscheinlichkeit r zu, bedeutet dies, daß wir vermuten, in einer langen Folge von (imaginären) zufälligen Ereignissen die e hinreichend ähnlich sind, trete die Eigenschaft ϕ mit einer relativen Häufigkeit r auf.

Auf der Basis dieser Interpretation von subjektiver Wahrscheinlichkeit läßt sich nunmehr erklären, auf welche Weise statistische Information zur Definition subjektiver Wahrscheinlichkeiten verwendet wird.

Die Grundvoraussetzung für die Benutzung statistischen Wissens zur Abschätzung einer relativen Häufigkeit in einer imaginären Folge von Ereignissen ist die Annahme, daß die Ereignisse in dieser Folge durch einen Zufallsprozeß zustande kommen, dessen statistische Verteilung durch die gegebene statistische Information beschrieben wird. Ein Beispiel mag dies erläutern: Gemäß unserer Interpretation bedeutet die oben angeführte subjektive Wahrscheinlichkeit von mindestens 0.7 dafür, daß ein gerade im Fernsehen laufender Film amerikanischen Ursprungs ist, daß wir erwarten, in einer langen Folge von Filmen, die wir zufällig im Fernsehen zu sehen bekommen und welche uns alle dieselben relevanten Indizien für ihre Herkunft liefern wie der tatsächlich beobachtete Film, mindestens einen Anteil von 70% amerikanischen Filmen zu finden. Wollen wir nun zusätzlich aus der Tatsache, daß 80% aller amerikanischen Kriminalfilme ein Happy End besitzen, folgern, daß in unserem Gedankenexperiment auch mindestens 56% Filme mit einem Happy End vorkommen müssen, setzt dies die Annahme voraus, daß die zufällige Beobachtung des Filmes im Fernsehen als zufällige Stichprobe gemäß der statistischen Verteilung gesehen werden kann – in diesem Beispiel also, daß jeder amerikanische Kriminalfilm mit gleicher Wahrscheinlichkeit der gerade gezeigte sein kann. Sollten wir etwa wissen, daß der Sender, welcher den Film ausstrahlt, gegenwärtig eine Sendereihe mit Filmen aus der "schwarzen Serie" zeigt, so werden wir die allgemeine statistische Information über alle amerikanischen

Kriminalfilme nicht zur Ermittlung von subjektiven Wahrscheinlichkeiten über den gerade zu sehenden heranziehen.

Auf Grund dieser und einiger weiterer Detailüberlegungen gelangen wir zu einem präzisen epistemischen Modell für "default reasoning about probabilities". Dieses epistemische Modell hinwiederum ist formalisierbar in Form eines statistischen Modells, in welchem die relativen Häufigkeiten von gewissen Eigenschaften in einer Folge von Ereignissen durch die empirischen Verteilungen in einer Folge von Zufallsvariablen repräsentiert werden. Für uns wichtige Aspekte des Verhaltens dieser empirischen Verteilungen sind in der Statistik in der Theorie der großen Abweichungen (large deviation) vom Erwartungswert untersucht worden. Die Resultate dieser Theorie erlauben uns zu zeigen, daß unser epistemisches Modell für die Kombination von statistischen mit subjektiven Wahrscheinlichkeiten das Prinzip der Cross-Entropy (Kullback-Leibler Abstand) Minimierung impliziert: Die subjektiven Wahrscheinlichkeiten, die wir auf Grund von a priori gegebenen subjektiven Wahrscheinlichkeiten und statistischen Wahrscheinlichkeiten ermitteln, werden durch das Wahrscheinlichkeitsmaß gegeben, welches innerhalb all jener Wahrscheinlichkeitsmaße, die mit der partiellen a priori Beschreibung des subjektiven Maßes konsistent sind, den minimalen Cross-Entropy Abstand zum statistischen Maß besitzt.

Nach dieser Herleitung des Cross-Entropy Minimierungs Prinzips auf Grund einer epistemischen Analyse, können wir daran gehen, das Schließen mit subjektiven Wahrscheinlichkeiten in unsere Logik zu integrieren. Zunächst muß hierfür noch ein Problem aus dem Wege geräumt werden: Wir haben Cross-Entropy Minimierung als den zentralen Prozeß für "default reasoning about probabilities" identifiziert. Die Cross-Entropy Funktion ist aber zunächst nur als Funktion auf reellwertigen Wahrscheinlichkeitsmaßen erklärt. In unsere Logik arbeiten wir jedoch mit Wahrscheinlichkeiten in reell abgeschlossenen Körpern. Wir müssen also untersuchen, ob sich Cross-Entropy in sinnvoller Weise auf solcherart verallgemeinerte Maße ausdehnen läßt.

In Kapitel 4 erweitern wir deshalb reell abgeschlossene Körper um eine Logarithmusfunktion, mit deren Hilfe dann eine verallgemeinerte Cross-Entropy Funktion definierbar wird. Es zeigt sich, daß sich die zentralen Eigenschaften des Cross-Entropy Minimierungsprozesses (insbesondere die von Shore und Johnson [1980] gezeigten Eigenschaften) allein auf der Basis der für die allgemeine Logarithmusfunktion gegebenen Axiome beweisen lassen, was die Verwendung der verallgemeinerten Cross-Entropy Funktion für "default reasoning about probabilities" mit Wahrscheinlichkeiten in reell abgeschlossenen Körpern rechtfertigt.

Der Ausbau der Logik $\mathscr{L}^\sigma$ zu einer Logik $\mathscr{L}^\beta$ zur kombinierten Repräsentation von statistischen und subjektiven Wahrscheinlichkeiten erfolgt in Kapitel 5. Während wir bei $\mathscr{L}^\sigma$ weitgehend den Definitionen von Bacchus [1990a] folgten, beschreiten wir hier andere Wege, als in vorangehenden Ansätzen ([Bacchus, 1990b], [Halpern, 1990]) gewählt wurden.

Die Syntax von L^σ erweitern wir um eine Konstruktionsmöglichkeit für Terme zur Repräsentation von subjektiven Wahrscheinlichkeiten. Wir gehen von der Annahme aus, daß die zu repräsentierenden subjektiven Wahrscheinlichkeiten eine Menge von Ereignissen $\mathbf{e}$ betreffen, über deren Wahrscheinlichkeit gewisse Eigenschaften ϕ zu besitzen, etwas ausgesagt werden soll. Ein subjektiver Wahrscheinlichkeitsterm hat dann die Form

$$\mathrm{prob}(\phi[\mathbf{e}]),$$

wobei ϕ eine Formel aus L^σ ist.

Interpretiert wird die um dieses Konstrukt erweiterte Sprache, die wir mit L^β bezeichnen, durch Strukturen, die wir aus einer statistischen Struktur $\mathfrak{M}$ gewinnen durch Hinzunahme eines weiteren Wahrscheinlichkeitsmaßes $\nu_{\mathbf{e}}$ auf der Algebra $\mathfrak{A}_n$, wobei n der Anzahl der Ereignissymbole in $\mathbf{e}$ bezeichnet. Die Interpretation eines subjektiven Wahrscheinlichkeitstermes $\mathrm{prob}(\phi[\mathbf{e}])$ in einer solchen Struktur $(\mathfrak{M},\nu_{\mathbf{e}})$ wird nunmehr gegeben durch

$$\nu_{\mathbf{e}}(\{\boldsymbol{a} \mid (\mathfrak{M}, \nu_{\mathbf{e}}) \models \phi[\boldsymbol{a}]\}).$$

Im Gegensatz zu den o.g. früheren Ansätzen werden subjektive Wahrscheinlichkeiten also nicht über einer Menge von möglichen Welten interpretiert und somit über einem ganz anderen Wahrscheinlichkeitsraum als die statistischen Wahrscheinlichkeiten, sondern gleichfalls vermöge eines Wahrscheinlichkeitsmaßes auf dem Träger. Auf diese Weise wird es möglich, Cross-Entropy Minimierung zwischen dem statistischen Maß μ_n und dem subjektiven Maß $\nu_{\mathbf{e}}$ durchzuführen. Dies dient als Grundlage, um für L^β den Begriff eines präferenziellen (oder default) Modells einzuführen: Eine Struktur $(\mathfrak{M},\nu_{\mathbf{e}})$ ist ein Default Modell von $\phi \in \mathrm{L}^\beta$, wenn kein $\nu'_{\mathbf{e}}$ existiert, derart daß die Struktur $(\mathfrak{M}, \nu'_{\mathbf{e}})$ gleichfalls ein Modell von ϕ ist und $\nu'_{\mathbf{e}}$ geringere Cross-Entropy zu $\mu_{\mathbf{e}}$ hat als $\nu_{\mathbf{e}}$. Indem wir uns auf Default Modelle beschränken, erhalten wir eine verschärfte Semantik für L^β, deren Folgerungsbeziehung $\models\!\!\!\approx$ "default reasoning about probabilities" formalisiert.

Abschließend (in Abschnitt 5.6) zeigen wir, daß sich die Eigenschaft, Default Modell zu sein, in L^β axiomatisieren läßt. Die Konsequenz hieraus ist, daß sich die Inferenzrelation $\models\!\!\!\approx$ auf L^β wiederum vollständig im Rahmen der Prädikatenlogik erster Stufe darstellen läßt. Insbesondere gibt es einen vollständigen logischen Kalkül für diese Inferenzrelation.

Introduction

1.

> “[The subject of probability theory] is the domain of mass phenomena and repetitive events, as the totality of spatial phenomena is the subject of geometry.”([von Mises, 1951])
>
> “With the usage just recommended, the term ‘frequency theory of probability’ is a pure incongruity; just as much so as ‘theory of square circles’.” ([Jaynes, 1978])

Like no other fundamental mathematical concept, the notion of probability has given rise to philosophical argument about its true meaning. While almost any two geometers will be in agreement to what real world phenomena their concept of “point” and “line” can rightfully be applied, and number theorists rarely engage in serious dispute about the meaning of a natural number, a similar understanding does not prevail among probability theorists.

The interpretation of probability that causes the dissent expressed in the two citations above is the frequentistic (also called objectivistic or empirical) interpretation. According to this view, probabilities can only be assigned to some property with reference to a large class of objects or events, each of which may or may not exhibit that property. Von Mises [1951] cites as examples the property of having flowers of a specific colour in a group of plants grown from a certain supply of seeds, the property of turning up heads in a long sequence of tosses of a coin, and the property of dying at a certain age in the group of all holders of a life insurance policy. In each of these examples, according to the frequentistic school of thought, we may speak of the probability of the given property (in the specified class). Whereas “Not only can we not say anything about the probability of death for a specific individual, however much may be known about him, but the expression itself does not have any meaning for us”([von Mises, 1951]).

To speak of probabilities of individual events, in contrast, is what the subjectivistic (personal, Bayesian) interpretation deems to be the only usage that accords with the true meaning of probability: “we recognize that a probability assignment is a means of describing a state of knowledge.”([Jaynes, 1978]). In [Savage, 1954], probably the most influential work promulgating the subjectivistic point of view, we find as examples of events to which probabilities may be assigned the event that a republican president will be elected in 1996, the event that it did snow in Chicago sometime in the month of May, 1995, or the event that a particular egg used for making an omelette is rotten.

2.

When it is our aim to design knowledge based systems that incorporate a broad section of a human reasoner's handling of uncertain and quantitative information, we are not in a position to take an extreme stance and declare one or another notion of probability as invalid or useless. Rather, we must acknowledge that in the real world we do encounter two distinguishable usages of the word probability.

> "The probability of being dealt a full house or better in a game of poker is less than 0.01." and
> "The probability that my opponent now has a full house or better is at least 0.2."

are two sentences that we would like a knowledge based system both to cope with, instead of rejecting the one or the other as a "syntax error", being a frequentistic or subjectivistic system, as the case may be. It is advisable, therefore, to adopt a pragmatic point of view, and just admit that probabilities come in two flavours: as the expression of a relative frequency in a large class of objects or events, and as a number expressing an uncertainty about an individual event.

Carnap was the first who admitted and studied both concepts of probability in their own right. In [1950] he denotes the two types of probability by probability$_1$ and probability$_2$, representing, respectively, a "degree of confirmation" and a "relative frequency in the long run".

This begs the question: should we not do away with the controversial term probability completely, and replace it by the expressions confirmation and frequency, or similar ones; thus avoiding the philosophical dispute and precluding possible misunderstandings? Two reasons stand against this: first, probability, after all, is the word that is commonly employed. The problem that this happens in several incongruous ways it is better to resolve by analyzing all usages, rather than to brush it aside by substituting new expressions. Secondly, different though their semantical content may be, all notions of probability give rise to essentially the same mathematical theory. Whether probabilities are seen as degrees of confirmation, or as relative frequencies, the rules that govern their manipulation are given by the same calculus, the authoritative axiomatization of which was given by Kolmogorov [1950].

It is not true, however, that all concepts of probability necessarily have identical mathematical properties. Kolmogorov actually provided two versions of his axiomatization, one in which probabilities are required to be only finitely additive, and one in which the axiom of countable additivity is added. The two versions of the additivity axiom have an unequal appeal to proponents of the different probability concepts. Von Mises [1951] proves that his frequency interpretation entails countable additivity, while Savage only assumes finite additivity: "I know of no argument leading to the requirement of countable additivity, and many of us have a strong intuitive tendency to regard as natural probability problems about the necessarily only finitely additive uniform densities on the integers,[...]" ([1954]).

That basically equivalent mathematical theories emanate from different concepts of probability, most likely is the reason that even after Carnap the study of various types of probability in parallel did not gain much impetus. It seems that only the emergence of human reasoning under uncertainty as a research topic in artificial intelligence caused a revival of interest in a (more or less) impartial investigation of the semantics of different kinds of probability statements.

The most comprehensive study in this field is the combined work of Halpern [1990] and Bacchus [1990b]. Here a twofold extension of first-order predicate logic is developed that integrates two types of probabilistic statements into a formal logical syntax, and provides semantics for the resulting language. Bacchus calls probabilities of one kind statistical, these correspond to relative frequencies, and those of the second kind propositional, because these are assigned to propositions in the language. Bacchus eschews the use of the term subjective probability for the latter ones. He does, however, view probabilities on propositions as an expression of a degree of belief. Since, as yet, the case has not convincingly been made that a degree of belief in a proposition – even for the most ideal rational agent – can be understood as resulting from an objective logical relationship between the available evidence and the proposition (as envisaged by Keynes [1921] and Carnap [1950]), it may be more to the point to make the seemingly inevitable subjectivity in such probability assignments transparent in our terminology. For this reason, we follow Bacchus only half the way, and, in the future, speak of *statistical probabilities* and *subjective probabilities.*

3.

Once it is realized that we have to deal with two separate kinds of probability when reasoning under uncertainty is to be formalized, we are faced with the question of how statistical and subjective probabilities are connected. It appears to be obvious that there is some interaction between the two concepts: when we meet a poker player who makes the two statements cited above, we immediately know that he must have made some observation that indicated his opponent to have a fairly good hand. Certainly, if the cards had just been dealt, previous to any reaction of the other player, a rational agent would not assign a probability as high as 0.2 to an event that belongs to a class with a much lower statistical probability. Rather, the statistical probability ≤ 0.01 for being dealt as good a hand as a full house would have served as an initial value for the subjective probability that the other player in the present situation is in possession of such a hand.

This is the principle of *direct (inductive) inference* that has been proposed both by Carnap [1950] and Reichenbach [1949]: if the relative frequency of a specific property (here: being a full house or better) in a *reference class* (here: the class of dealings of five cards from a shuffled deck) is r, and the evidence obtained about a specific object or event implies that it belongs to that reference class, then r is the subjective probability that should rationally be assigned to the proposition that the specific object or event has the given property. In most situations, of course, the problem will arise that the evidence establishes membership in more than one reference class. To find rules for choosing the most appropriate reference class is the main difficulty that has to be conquered in order to make systematic use of direct inference.

An assignment of a subjective probability by direct inference is subject to revision when further evidence is obtained, because such evidence may persuade us to consider a new reference class as being the most appropriate one. This makes the direct inference principle share a key formal aspect with *default* or *nonmonotonic logic* as pioneered, among others, by McCarthy [1980] and Reiter [1980]: like these, it is nonmonotonic, i.e. additional information can invalidate previous inferences. The analogy between a direct inference system and nonmonotonic logic, however, is not limited to this one formal property. They also are similar in that they are designed to formalize quite similar forms of commonsense reasoning. Nonmonotonic logic formalizes inferences which lead us to assume that a specific object has a specific property, when this property typically holds for objects from the domain under consideration, and the given object is not known to be exceptional by not possessing that property. "Typically" here has to be understood in such a strong sense, that the set of objects that are exceptional with respect to the property in question is negligible compared to the set of objects that are normal with respect to the property. Deriving subjective probabilities by direct inference can be seen as a quantitative analogue to this kind of qualitative default reasoning. Where in the latter knowledge of typicality of a property is used as input to derive (defeasible, but for the time being 0,1-valued) default conclusions, in the former specific quantitative information is used to derive probabilistic default conclusions. Because of this analogy, we use the expression *default reasoning about probabilities* for inferences that use statistical information to derive subjective probabilities.

4.

Direct inference only is applicable when evidence has been obtained that definitely establishes membership of the object or event under consideration in a reference class for which statistical data is available. This ideal situation will not always be encountered in reality.

Consider, for instance, a mechanic who wants to repair a car whose engine will not start. On turning the key the starter does turn, but gives off a somewhat weaker sound than the mechanic would expect from a properly functioning starter. This indicates that the problem might be caused by the battery being low. In this case it can perhaps be fixed simply by recharging the battery. It is also possible that the battery must be replaced. It is likely that the mechanic has some reliable statistics about the percentage of cases of low batteries in which it is sufficient to charge the battery. On the other hand, it is not very likely that she would feel able to immediately state a statistical probability for a charging of the battery being a sufficient remedy for the start up problem, given the specific sound made by the starter. In order to estimate the probability that in the case at hand recharging will do, the mechanic will therefore employ a two stage process: first she will use the evidence of the starter's sound to assign a subjective probability for the trouble being caused by a low battery, then the statistical information is employed to derive the specific probability that the old battery need not be replaced.

For the mechanic to first transform the actual evidence into subjective probabilities for some propositions, before statistical knowledge can be brought to bear on the deduction of further subjective probabilities, becomes inevitable when she uses an electronic decision support system for assisting her diagnoses. Such a system, for the foreseeable future, will only be able to process information encoded in a certain more or less restricted formal language. In that language it will be possible to express that battery power is low, or that this is true with a certain probability – it will probably not be possible to directly feed the sound made by the starter into the system. The system then has to be able to make use of statistical information in its data base and subjective probabilities supplied by the user in order to attach probabilities to various possible diagnoses.

We conclude that default reasoning about probabilities can not be limited to combining deterministic evidence with statistical knowledge by some scheme of direct inference. Default reasoning about probabilities more often takes the form of combining statistical data with some given subjective probabilities to obtain a more complete set of subjective probabilities. This is true for a human reasoner because he or she needs to structure the inference of a subjective probability using other subjective probabilities as intermediate results; for an implementation of default reasoning about probabilities in a knowledge based system such a combination is furthermore mandated by the restrictiveness of the representation language used by the system, which will not permit the user to directly enter every piece of deterministic, observed evidence.

5.

To define default reasoning about probabilities as a combination of statistical data with the partial description of a subjective probability distribution makes this process formally equivalent to the process of *updating*.

Inference problems that ask for an update of a probability distribution on the basis of newly obtained evidence occur in fairly diverse contexts. For example we may want to update an estimate for the parameters of a (statistical) probability distribution governing a physical system when new measurements have been made of the system. A subjective probability distribution describing an agent's state of knowledge may have to be updated when new information is given to the agent.

When we view a statistical probability measure as the prior subjective probability measure before any evidence about an object or event has been obtained, we can view default reasoning about probabilities as a special case of updating. It is therefore not surprising that methods that have been proposed for updating probability measures are interesting candidates for solving problems of default reasoning about probability that are beyond the scope of direct inference.

The probably most widely accepted rule for updating has originally been proposed by Jeffrey [1965] for the context of subjective probabilities. It applies in the situation where the evidence provides the posterior probability values $q(A_i)$ for a finite number of mutually disjoint and exhaustive subsets $A_1, \ldots, A_n$ of the domain of possible events. *Jeffrey's rule* then states that the complete posterior distribution q should be defined by retaining the conditional probabilities of the prior distribution p on each of the A_i, i.e. for each set of events A:

$$q(A) = \sum_{i=1}^{n} q(A_i) p(A \mid A_i).$$

Interpreting p as a statistical measure and q as a subjective probability measure, this can be read as a rule for default reasoning about probabilities.

In our previous example, supposing the statistical probability of fixing a start-up problem by recharging the battery is 0.3 when the problem is caused by a low battery, and 0 otherwise, furthermore assuming that the mechanic, on the basis of the sound made by the starter, assigns a subjective probability of 0.6 to A_1: "the problem is caused by a low battery", we can apply Jeffrey's rule to A_1, A_2: "not A_1", and A: "the problem can be fixed by recharging the battery", to obtain a subjective probability $q(A) = 0.6 \cdot 0.3 + 0.4 \cdot 0 = 0.18$.

Jeffrey's rule is very intuitive, but, unfortunately, still rather limited in its applicability. More general rules for updating therefore have been sought, rules that can also be applied when the sets A_i to which the given information on the posterior distribution refers are not assumed to be mutually disjoint, and the information does not necessarily prescribe exact probability values for these sets. In cases where Jeffrey's rule is applicable such a more general rule, of course, should yield the same result as Jeffrey's rule.

Among rules that satisfy these desiderata, updating according to the *minimum cross-entropy principle* has received the greatest amount of attention. By this rule, from the set of probability measures that are consistent with the given information, that measure is chosen as the posterior q that minimizes the cross-entropy function

$$\sum_e q(e)\ln(q(e)/p(e))$$

with e ranging over the (at most countable) set of possible events. Again, this rule can just as well be read as a rule for default reasoning about probabilities.

Lacking the immediate, intuitive appeal of Jeffrey's rule, the use of the minimum cross-entropy principle has to be justified by somewhat more sophisticated arguments than had to be advanced in favour of the former. Shore and Johnson [1980] present such an argument, showing that cross-entropy minimization is the only updating procedure that satisfies a certain set of intuitive axioms.

Overview

In the present work a logical formalism based on first-order predicate logic is developed for reasoning both about statistical and subjective probabilities.

We begin with an examination of a logic for reasoning with statistical probabilities only. This logic is defined very similarly as the one given by Bacchus [1990a]. Particularly, probability values will not be required to always be real numbers, but may be taken from any real-closed field. As in [Bacchus, 1990a], this permits us to derive a completeness result for that logic. The derivation of that result given here, however, is of a somewhat different nature than the one presented by Bacchus: it will be shown that the given probabilistic extension of first-order logic can be encoded in first-order logic itself (and therefore is not really a true extension). The completeness of first-order logic then entails the completeness of the probabilistic logic.

A central part of this work consists of the derivation of the minimum cross-entropy principle for default reasoning about probabilities in chapter 3. Here a new interpretation of the meaning of subjective probabilities is introduced from which an epistemic model is derived for how default reasoning about probabilities is actually performed. The epistemic model then, in turn, is formalized in a statistical model in which the minimum cross-entropy principle can be derived.

Having thus established cross-entropy minimization as the adequate analytical tool for modeling default reasoning about probabilities, we extend the representation language for statistical probabilities by introducing expressions representing subjective probabilities. Deviating from previous approaches ([Halpern, 1990],[Bacchus, 1990b]), where subjective probability expressions are interpreted by a probability distribution over possible worlds, we shall use a probability measure on the domain for that purpose. Interpreting statistical and subjective probabilities by measures on the same probability space enables us to define a preferred model semantics for the language that implements the minimum cross-entropy principle, so that entailment with respect to this semantics is a formal model for default reasoning about probabilities. The resulting logic is shown to be completely representable within first-order predicate logic, so that we receive a system in which probabilistic default inferences are conducted on the level of the representation language itself, where in previous approaches ([Bacchus *et al.*, 1992], [Paris and Vencovská, 1992]) always an extra-logical inference mechanism has been used. Particularly, in this manner, we obtain a completeness result for our system for default reasoning about probabilities.

To conclude, it might be helpful to state explicitly two topics that this work is not about.

First, we are here not concerned with making default assumptions about statistical distributions. In many cases the available statistical information will not specify a unique probability distribution. One may be tempted in these cases to also subject the statistical information to some inference process that selects one of the probability measures consistent with the given data as the most reasonable guess for the true statistical distribution. However, it is much harder to justify such an inference process for statistical probabilities than for subjective probabilities: for the latter ones we know that a human reasoner actually employs some kinds of (default) reasoning processes to define his or her subjective probabilities. Statistical probabilities, on the other hand, describe objective properties of the world. To substitute default assumptions for missing information about the true statistical distribution requires that we have some meta-level knowledge about which statistical distributions are more likely than others on the domain under consideration (knowledge as might be obtained, for example, by drawing a random sample of elements from the domain, an analysis of which can lead to an estimate for the true distribution). Since such meta-level knowledge can not be represented in the formal language we use, and there is no reason to believe that relative likelihoods of statistical distributions are determined by the same rule for every domain of objects or events that might be described in our language, we here refrain from doing any default inferencing of statistical probabilities.

The second topic that we will not be concerned with is probabilistic semantics for logical default reasoning. The close analogy between the statement of a logical default "typically an A is a B", and a probabilistic statement "with a high probability an A is a B" has led to several proposals to interpret logical default reasoning by probabilistic semantics ([Pearl, 1989], [Bacchus *et al.*, 1993]). However, it seems that somewhat different questions arise when probabilistic reasoning is examined as a basis for logical default reasoning, in which case the main concern lies with extreme, i.e. "almost" 0,1-valued probabilities, than when it is examined in view of application to non-extreme probability values, which is the subject of the present work.

To forestall a possible confusion between default reasoning about probabilities as treated here, and logical default reasoning on the basis of a probabilistic interpretation, has been the reason for choosing the somewhat clumsy expression to designate the former. The term "probabilistic default reasoning" would have made for smoother reading, but might be considered a little ambiguous.

As a further measure to mark the distinctness of our subject from nonmonotonic logic, all references to ornithological questions, specifically the flying abilities of members of the order sphenisciformes, have been purged from this work.

Chapter 1

Preliminaries

1.1 Notation

Most of the mathematical notation used in this work is standard and requires no special introduction. A few conventions used should be pointed out explicitly, however.

The sets of natural, rational, and real numbers are respectively designated by $\mathbf{N}$, $\mathbf{Q}$, and $\mathbf{R}$.

The symbol $\subseteq$ is used for the subset relation, while $\subset$ is reserved to denote the strict subset relation. Analogously for $\supseteq$ and $\supset$. Disjointness of a union is represented by $\dot{\cup}$. A^c stands for the complement of the set A. The restriction of a function f or relation R to a subset A of its domain is denoted by $f\restriction A$ ($R\restriction A$).

An indexed family (especially: a sequence) $\{A_i \mid i \in I\}$ of mathematical objects A_i of some kind is denoted by $(A_i)_{i\in I}$. This notation is particularly useful when either I, or both i and I are clear in a given context, in which cases we simply write $(A_i)_i$ or (A_i), respectively.

Some special attention we have to devote to notation dealing with *tuples*, which we here introduce in some detail. For I a finite subset of $\mathbf{N}$, and A_i $(i \in I)$ a family of sets,

$$\times_{i\in I} A_i = \{f : I \to \cup_{i\in I} A_i \mid f(i) \in A_i\}$$

is the *cartesian product* of the A_i. Its elements we denote by boldface characters $\boldsymbol{a}$, $\boldsymbol{b}$,... If $A_i = A$ for all $i \in I$, then we write A^I for $\times_{i\in I} A_i$. Elements of A^I are called I-tuples of elements of A. When $I = \{1, \ldots, n\}$, then A^I is the set of n-tuples of elements of A and also denoted by A^n. An n-tuple $\boldsymbol{a}$ with $\boldsymbol{a}(i)=a_i$ may be written in the form $(a_1, \ldots, a_n)$. We use $|\,\boldsymbol{a}\,|$ for the *length* of the tuple $\boldsymbol{a} \in A^I$, i.e. the cardinality of I.

By a slight abuse of notation, we occasionally identify a tuple $\boldsymbol{a} \in A^I$ with the set of its components $\{\boldsymbol{a}(i) \mid i \in I\}$, and use expressions like $a \in \boldsymbol{a}$ or $\boldsymbol{a} \cup \boldsymbol{b}$. For $f : A \to B$, $\boldsymbol{a} \in A^I$ and $\boldsymbol{b} \in B^I$ we write $f(\boldsymbol{a}) = \boldsymbol{b}$ when $f(\boldsymbol{a}(i)) = \boldsymbol{b}(i)$ for all $i \in I$.

When $I \cap I' = \emptyset$, we may identify $A^I \times A^{I'}$ with $A^{I\cup I'}$. For $\boldsymbol{a} \in A^I$, $\boldsymbol{b} \in A^{I'}$ then $(\boldsymbol{a},\boldsymbol{b})$ denotes the element $\boldsymbol{c}$ of $A^{I\cup I'}$ with $\boldsymbol{c}\restriction I = \boldsymbol{a}$ and $\boldsymbol{c}\restriction I' = \boldsymbol{b}$.

For a permutation π of I (i.e., a bijection of I), and $\boldsymbol{a} \in A^I$, we write $\pi\boldsymbol{a}$ for the element $\boldsymbol{a} \circ \pi \in A^I$: $(\pi\boldsymbol{a})(i) = \boldsymbol{a}(\pi(i))$.

Model theoretic structures are denoted by old German capital letters $\mathfrak{A}$, $\mathfrak{F}$, $\mathfrak{M}$, ... The domain of such a structure is designated by the corresponding Roman letter (for the special case of structures that are fields of numbers: boldface Roman letter), and the interpretation of some function or relation symbol in that structure by the symbol with an adequate superscript. The structure of the real numbers as an ordered field, for instance, is represented as $\mathfrak{R} = (\mathbf{R}, +^{\mathrm{R}}, \cdot^{\mathrm{R}}, 0^{\mathrm{R}}, 1^{\mathrm{R}}, \leq^{\mathrm{R}})$.

In addition to the usual convention in logic to denote by $\phi(\boldsymbol{v})$ an expression whose free variables are among the variables $\boldsymbol{v}$, we use the denotation $\phi\langle\boldsymbol{v}\rangle$ for an expression with exactly the free variables $\boldsymbol{v}$. Writing $\phi(\boldsymbol{v},\boldsymbol{w})$ or $\phi\langle\boldsymbol{v},\boldsymbol{w}\rangle$ always is meant to imply that $\boldsymbol{v} \cap \boldsymbol{w} = \emptyset$.

Finally, to designate identity between syntactic objects (i.e. terms and formulas), we use the symbol $\equiv$ in order to avoid any confusion with the identity symbol $=$ used on the level of formal syntax.

1.2 Algebras and Measures

We give a concise review of the basic theory of finitely additive measures, which is an indispensable prerequisite for the entire work to follow.

Standard texts on measure theory (e.g. [Halmos, 1950], [Cohn, 1993]) being usually almost completely preoccupied with countably additive measures, systematic accounts of finitely additive measures are hard to come by. This obliges us to also give proofs for a couple of elementary statements that one might have expected to be easily retrievable from the literature.

Definition 1.2.1 Let M be a set, $\mathfrak{A} \subseteq 2^M$. $\mathfrak{A}$ is an *algebra over* M, iff $M \in \mathfrak{A}$, and $\mathfrak{A}$ is closed under finite unions and complements. If $\mathfrak{A}$ also is closed under countable unions, then $\mathfrak{A}$ is called a *$\sigma-$algebra*.

Thus, an algebra is just a boolean algebra of sets. Borrowing some terminology from the theory of boolean algebras, we call an algebra $\mathfrak{A}$ *atomic* iff for every $A \in \mathfrak{A}$ there exists an $A' \subseteq A$ so that $B \subset A'$ with $B \in \mathfrak{A}$ only holds for $B = \emptyset$. Elements A' of $\mathfrak{A}$ with this property are called the *atoms* of $\mathfrak{A}$. Observe that every finite algebra is atomic.

$\mathfrak{A}'$ is a *subalgebra* of $\mathfrak{A}$, if $\mathfrak{A}'$ is a subset of $\mathfrak{A}$ with $M \in \mathfrak{A}'$, and is itself an algebra.

If $\mathfrak{A}$ is an algebra, $A \in \mathfrak{A}$, then $\{B \cap A \mid B \in \mathfrak{A}\}$ is an algebra over A: the *relative algebra of* $\mathfrak{A}$ *with respect to* A.

For $\mathfrak{E} \subseteq 2^M$, $\mathfrak{A}(\mathfrak{E})$ denotes the smallest algebra containing $\mathfrak{E}$: the algebra *generated* by $\mathfrak{E}$. If $\mathfrak{E} = \{E_1, \ldots, E_n\}$, then $\mathfrak{A}(\mathfrak{E})$ is finite with atoms the nonempty elements of $\{\tilde{E}_1 \cap \ldots \cap \tilde{E}_n \mid \tilde{E}_i \in \{E_i, E_i^c\}\}$.

Example 1.2.2 Let M be infinite. Define

$$\mathfrak{A}^{\mathrm{c/f}} := \{A \subseteq M \mid A \text{ finite or } A^c \text{ finite}\}.$$

Then $\mathfrak{A}^{\mathrm{c/f}}$ is an algebra: the algebra of finite and co-finite subsets of M. $\mathfrak{A}^{\mathrm{c/f}}$ is atomic with the singleton sets $\{a\}$ $(a \in M)$ as atoms. The set of singletons also is a generating system for $\mathfrak{A}^{\mathrm{c/f}}$. $\mathfrak{A}^{\mathrm{c/f}}$ is a standard example in measure theory of an algebra that is not a $\sigma-$algebra, and a useful provider of counterexamples.

In standard measure theory, measures on a ($\sigma-$)algebra are defined to take values in the real numbers. Since we will later want to deal with measures in the formal framework of first-order predicate logic where the reals are not axiomatizable, we will here relax this definition somewhat, and allow measures to take values in arbitrary real closed fields.

Definition 1.2.3 Let $\mathrm{S}_{\mathrm{OF}} = \{0, 1, +, \cdot, \leq\}$ be the vocabulary of ordered fields. An S_{OF}-structure $\mathfrak{F}$=$\{\mathbf{F}, 0^{\mathrm{F}}, 1^{\mathrm{F}}, +^{\mathrm{F}}, \cdot^{\mathrm{F}}, \leq^{\mathrm{F}}\}$ is a *real closed field* (rc-field for short), if it satisfies the axioms RCF consisting of

- The field axioms.
- Axioms stating that $\leq$ is a total order.
- Two axioms for the compatibility of $\leq$ with the algebraic operations:

$$\forall xyz((x \leq y \wedge 0 \leq z) \rightarrow x \cdot z \leq y \cdot z),$$
$$\forall xyz(x \leq y \rightarrow x + z \leq y + z).$$

- An axiom for the existence of square roots

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

- A schema demanding that every polynomial of uneven degree has a root

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

Example 1.2.4 (a) The field $\mathfrak{F}_{\mathrm{al}}$ of algebraic numbers is an rc-field. $\mathfrak{F}_{\mathrm{al}}$ is a very special model of RCF: every rc-field $\mathfrak{F}$ contains a substructure $\mathfrak{F}'$ such that $\mathfrak{F}'$ is isomorphic to $\mathfrak{F}_{\mathrm{al}}$, and the elements of $\mathbf{F}'$ satisfy the same first-order formulas when considered inside $\mathfrak{F}'$ and when considered inside $\mathfrak{F}$. In short, every rc-field is an *elementary extension* of $\mathfrak{F}_{\mathrm{al}}$, or $\mathfrak{F}_{\mathrm{al}}$ is the *prime model* of RCF.

(b) The field of real numbers $\mathfrak{R}$ is an rc-field.

(c) Let $\mathfrak{R}^*$ be an elementary extension of $\mathfrak{R}$. By definition, $\mathfrak{R}^*$ and $\mathfrak{R}$ then are elementarily equivalent, so that $\mathfrak{R}^*$ is an rc-field. If $\mathfrak{R}^*$ is a proper extension, then $\mathbf{R}^*$ contains *infinitesimals*, i.e. elements $r^* \in \mathbf{R}^*$ with $0 < r^* < r$ for all $r \in \mathbf{R}$.

From the existence of a prime model postulated in part (a) of this example it follows that the theory of real closed fields is complete. $\mathfrak{R}$ being an rc-field, RCF thereby provides a recursively enumerable axiomatization of the S_{OF}-theory of $\mathbf{R}$ (see [Rabin, 1977] for details).

Definition 1.2.5 Let $\mathfrak{A}$ be an algebra over M, $\mathfrak{F}$ a real closed field. Let $\mathbf{F}^+ := \{x \in \mathbf{F} \mid 0 \leq x\}$. A function

$$\mu : \mathfrak{A} \rightarrow \mathbf{F}^+$$

is a *measure* iff $\mu(\emptyset) = 0$, and $\mu(A \cup B) = \mu(A) + \mu(B)$ for all $A, B \in \mathfrak{A}$ with $A \cap B = \emptyset$. μ is a *probability measure* iff $\mu(M) = 1$. A pair $(\mathfrak{A}, \mu)$ with $\mathfrak{A}$ an algebra and μ a (probability) measure on $\mathfrak{A}$ is called a *(probability) measure algebra.*

If $\mathfrak{A}$ is a $\sigma-$algebra, $\mathbf{F}$=$\mathbf{R}$, and μ is σ-additive, i.e. $\mu(\cup_{i\geq 1} A_i) = \sum_{i\geq 1} \mu(A_i)$ for $\{A_1, A_2, \ldots\}$ a countable family of mutually disjoint sets, then μ is called a (probability) σ-measure.

Definition 1.2.6 The set of all probability measures with values in $\mathfrak{F}$ on the algebra $\mathfrak{A}$ is denoted by

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

We write simply $\Delta\mathfrak{A}$ for $\Delta_{\mathrm{R}}\mathfrak{A}$.

When $\mathfrak{A}$ is a finite algebra with atoms $A_1, \ldots, A_N$, we can identify $\Delta_{\mathrm{F}}\mathfrak{A}$ with

$$\Delta^N_{\mathrm{F}} := \{(r_1, \ldots, r_N) \in \mathbf{F}^N \mid r_i \geq 0, \ \sum r_i = 1\}.$$

For Δ^N_{R} we simply write Δ^N. For $\boldsymbol{r}$,$\boldsymbol{s} \in \Delta^N_{\mathrm{F}}$ the Euclidean distance

$$|\, \boldsymbol{r} - \boldsymbol{s}\,| := \left(\sum (\boldsymbol{r}(i) - \boldsymbol{s}(i))^2\right)^{\frac{1}{2}}$$

is defined, which induces a topology on Δ^N_{F}. When standard topological concepts like openness of a set, the closure $cl\,A$, interior $int\,A$, or boundary $bd\,A$ of a set A are employed for subsets of Δ^N_{F}, it is always with respect to this topology.

The i-th component of a measure $\mu \in \Delta^N_{\mathrm{F}}$ usually is denoted by μ_i, i.e. $\mu = (\mu_1, \ldots, \mu_N)$. Unfortunately, this provokes a certain shortage of denotation, since we will also have to use $\mu_1, \mu_2, \ldots$ for members of an indexed family of probability measures. Hence, μ_i can either be a probability measure itself, or the i-th component of the probability measure μ (it does not help to use superscripts μ^i in one of the two cases, because this denotation will be used for product measure, see below). In context, however, it should always be sufficiently clear, which of the two readings of μ_i is the intended one.

One remark on the terminology here should be added: in measure theory in general, and probability theory in particular, one is predominantly concerned with σ-measures, which therefore are usually simply called (probability-) measures. A finitely additive measure sometimes is called a content. Since we, on the other hand, will almost exclusively deal with finitely additive measures, we reserve the most simple and intuitive term for these. Also, we sometimes will be using the term *probability distribution* as a synonym for probability measure.

Example 1.2.7 Let M be a set, $a_1, a_2, \ldots \in M$, $p_1, p_2, \ldots \in \mathbf{R}$ with $p_i \geq 0$ and $\sum p_i = 1$. For $A \subseteq M$ let

$$\mu(A) := \sum_{a_i \in A} p_i.$$

Then μ is a probability σ-measure on the σ-algebra 2^M. Measures defined in this way we call *real discrete measures.* Note that in arbitrary rc-fields the infinite sum $\sum p_i$ need not be defined, so that we really have to assume real-valued probabilities, unless all but finitely many of the p_i are zero.

Example 1.2.8 Let $\mathbf{F}$ be an rc-field, $\mathfrak{A}^{\mathrm{c/f}}$ as in example 1.2.2. For $A \in \mathfrak{A}^{\mathrm{c/f}}$ define

$$\mu^{\mathrm{c/f}}(A) := \begin{cases} 0^{\mathrm{F}} & A \text{ finite} \\ 1^{\mathrm{F}} & A^c \text{ finite.} \end{cases}$$

Then $\mu^{\mathrm{c/f}}$ is a probability measure on $\mathfrak{A}^{\mathrm{c/f}}$. Measures that, like $\mu^{\mathrm{c/f}}$, assign zero probability to every singleton set we call *continuous* (because for the special case of probability measures on $\mathbf{R}$, these are just the measures that have a continuous distribution function).

Generating systems $\mathfrak{E}$ for an algebra $\mathfrak{A}$ are of particular interest, when a given additive function

$$\mu_0 : \ \mathfrak{E} \to \mathbf{F}^+$$

already uniquely determines a measure μ on $\mathfrak{A}$. This is the case when $\mathfrak{E}$ disjointly generates $\mathfrak{A}(\mathfrak{E})$ in the sense of the following definition.

Definition 1.2.9 $\mathfrak{E}$ *disjointly generates* $\mathfrak{A}$ iff every $A \in \mathfrak{A}$ is the disjoint finite union of elements of $\mathfrak{E}$.

Theorem 1.2.10 Let $\mathfrak{A}$ be disjointly generated by $\mathfrak{E}$, $\mu_0 : \ \mathfrak{E} \to \mathbf{F}^+$ be finitely additive. The unique measure μ on $\mathfrak{A}$ extending μ_0 is defined by

$$\mu(A) := \sum_{i=1}^{n} \mu_0(E_i) \tag{1.1}$$

$A \in \mathfrak{A}$ the disjoint union of $E_1, \ldots, E_n \in \mathfrak{E}$.

Proof: Obviously, every measure on $\mathfrak{A}$ extending μ_0 must satisfy (1.1), so there can be at most one such measure. To show that (1.1) defines a measure, we only have to show that the definition of $\mu(A)$ is independent of the particular representation $\dot\bigcup_{i=1,\ldots,n} E_i$ for A. For this purpose, let $\dot\bigcup_{j=1,\ldots,m} F_j$ be an alternative representation of A. For each pair $(i,j) \in \{1,\ldots,n\} \times \{1,\ldots,m\}$ there exist $H_k^{(i,j)} \in \mathfrak{E}$ $(k = 1,\ldots,l_{(i,j)})$ with $E_i \cap F_j = \dot\bigcup_{k=1,\ldots,l_{(i,j)}} H_k^{(i,j)}$. Then

$$\sum_{i=1}^{n} \mu_0(E_i) = \sum_{i=1}^{n} \sum_{\substack{j=1,\ldots,m \\ k=1,\ldots,l_{(i,j)}}} \mu_0(H_k^{(i,j)}) = \sum_{j=1}^{m} \sum_{\substack{i=1,\ldots,n \\ k=1,\ldots,l_{(i,j)}}} \mu_0(H_k^{(i,j)}) = \sum_{j=1}^{m} \mu_0(F_j).$$

□

Definition 1.2.11 Let $\mathfrak{A}$,$\mathfrak{B}$ be algebras over sets M and N respectively. The *product algebra* $\mathfrak{A} \times \mathfrak{B}$ is the algebra over $M \times N$ generated by the system of *measurable rectangles*

$$\mathfrak{E}^{\times} := \{A \times B \mid A \in \mathfrak{A}, \ B \in \mathfrak{B}\}.$$

It is well known that $\mathfrak{E}^\times$ disjointly generates $\mathfrak{A} \times \mathfrak{B}$, and that for measures μ and ν on $\mathfrak{A}$ and $\mathfrak{B}$ respectively the function

$$\lambda(A \times B) := \mu(A)\nu(B)$$

is additive (see [Halmos, 1950] for instance). The unique measure extending λ to $\mathfrak{A} \times \mathfrak{B}$ is called the *product measure* of μ and ν, denoted by $\mu \otimes \nu$.

These definitions extend to the product of n algebras in a natural way. The product $\mathfrak{A} \times \ldots \times \mathfrak{A}$ (n factors) is denoted $\mathfrak{A}^n$. For the measure $\mu \otimes \ldots \otimes \mu$ we write μ^n.

Definition 1.2.12 Let μ be a measure on the product algebra $\mathfrak{A} \times \mathfrak{B}$ over $M \times N$. The *marginal distribution* $\mu \mid_1 \mathfrak{A}$ of μ on $\mathfrak{A}$ as the first component is defined by

$$(\mu \mid_1 \mathfrak{A})(A) = \mu(A \times N) \quad (A \in \mathfrak{A}).$$

The notation introduced in this definition at a first glance may seem somewhat redundant. However, we can not dispense with either the explicit reference to the algebra $\mathfrak{A}$ on which the marginal distribution is defined, nor the place it has in the original product: only writing $\mu \mid \mathfrak{A}$ would be ambiguous in the case where $\mathfrak{B} = \mathfrak{A}$; just writing $\mu \mid_1$ is ambiguous when the product in fact has more than two components, for example $\mathfrak{A} \times \mathfrak{B} = \mathfrak{A}_1 \times \mathfrak{A}_2 \times \mathfrak{B}$. Here $\mu \mid_1$ might refer to the marginal distribution on $\mathfrak{A}_1$, or on $\mathfrak{A}_1 \times \mathfrak{A}_2$.

We conclude this section by proving one rather specialized theorem that will only be needed at a particular point in section 2.4. Being of a purely algebraic nature, it nevertheless is best placed in the present context.

Theorem 1.2.13 Let $\mathfrak{A}$ be an algebra disjointly generated by $\mathfrak{E}$. Let $\mathfrak{B}$ be a finite algebra over the same set M as $\mathfrak{A}$, $B_1, \ldots, B_n$ the atoms of $\mathfrak{B}$. The algebra $\mathfrak{A}(\mathfrak{A} \cup \mathfrak{B})$ generated by $\mathfrak{A}$ and $\mathfrak{B}$ is disjointly generated by

$$\mathfrak{G} := \{E \cap B_i \mid E \in \mathfrak{E},\ i = 1, \ldots, n\}.$$

Proof: Let $\mathfrak{G}^{\dot\cup}$ denote the set of disjoint finite unions of elements of $\mathfrak{G}$. Clearly then

$$\mathfrak{A} \cup \mathfrak{B} \subseteq \mathfrak{G}^{\dot\cup} \subseteq \mathfrak{A}(\mathfrak{A} \cup \mathfrak{B}).$$

To prove the theorem, we therefore have to show that $\mathfrak{G}^{\dot\cup}$ is an algebra over M.

To show that $\mathfrak{G}^{\dot\cup}$ is closed under unions, let

$$G_1 = \dot{\bigcup_{j=1,\ldots,m}} (E^1_j \cap B_{i_j}), \quad G_2 = \dot{\bigcup_{k=1,\ldots,l}} (E^2_k \cap B_{i_k}) \ \in \mathfrak{G}^{\dot\cup}.$$

To show that $G_1 \cup G_2 \in \mathfrak{G}^{\dot\cup}$ it suffices in fact to only consider the case $l = 1$, so that we may let $G_2 = E^2 \cap B^2$ for some $E^2 \in \mathfrak{E},\ B^2 \in \{B_1, \ldots, B_n\}$. Then

$$\left[\dot{\bigcup_{j=1,\ldots,m}} (E^1_j \cap B_{i_j})\right] \cup (E^2 \cap B^2) \ = \ \dot{\bigcup_{\{j \mid B_{i_j} \neq B^2\}}} (E^1_j \cap B_{i_j}) \mathbin{\dot\cup} \left[\Big(\bigcup_{\{j \mid B_{i_j} = B^2\}} E^1_j \cup E^2\Big) \cap B^2\right].$$

The union $\cup_{\{j|B_{i_j}=B^2\}} E_j^1 \cup E^2$ is in $\mathfrak{A}$ and hence the disjoint union of suitable $F_h \in \mathfrak{E}$ ($h = 1, \ldots, p$). Thus,

$$(\bigcup_{\{j|B_{i_j}=B^2\}} E_j^1 \cup E^2) \cap B^2 = (\dot{\bigcup_{h=1,\ldots,p}} F_h^j) \cap B^2 = \dot{\bigcup_h} (F_h \cap B^2),$$

and we have obtained a representation showing that $G_1 \cup G_2$ is in $\mathfrak{G}^{\dot\cup}$. It remains to show that $\mathfrak{G}^{\dot\cup}$ is closed under complements. Let $G = \dot\cup_{j=1,\ldots,m}(E_j \cap B_{i_j}) \in \mathfrak{G}^{\dot\cup}$. Since

$$G^c = \dot{\bigcup_{i=1,\ldots,n}} (G^c \cap B_i),$$

it suffices to show that $G^c \cap B \in \mathfrak{G}^{\dot\cup}$ for fixed $B \in \{B_1, \ldots, B_n\}$. We have

$$[\dot{\bigcup_{j=1,\ldots,m}} (E_j \cap B_{i_j})]^c \cap B = [\dot{\bigcup_{\substack{j=1,\ldots,m \\ B_{i_j}=B}}} (E_j \cap B_{i_j})]^c \cap B,$$

so that the problem reduces to showing that sets of the form

$$[\dot{\bigcup_{j=1,\ldots,m}} (E_j \cap B)]^c \cap B$$

belong to $\mathfrak{G}^{\dot\cup}$:

$$\begin{aligned} [\dot{\bigcup_{j=1,\ldots,m}} (E_j \cap B)]^c \cap B &= [(\dot{\bigcup_{j=1,\ldots,m}} E_j)^c \cup B^c] \cap B \\ &= (\dot{\bigcup_{j=1,\ldots,m}} E_j)^c \cap B \\ &= (\dot{\bigcup_{h=1,\ldots,p}} F_h) \cap B \\ &= \dot{\bigcup_{h=1,\ldots,p}} (F_h \cap B) \quad \in \mathfrak{G}^{\dot\cup} \end{aligned}$$

with $\dot\cup_{h=1,\ldots,p} F_h$ a representation of $(\dot\cup_{j=1,\ldots,m} E_j)^c$ as the disjoint union of elements $F_h \in \mathfrak{E}$. □

Chapter 2

The Logic of Statistical Probabilities

2.1 Statistical Probabilities

Consider the statements

> "The probability of rolling a one with a throw of a die is 1/6"
> "Mystery films have a happy end with probability 0.7"
> "The probability that the child of an actor will become an actor is at least 0.1"

In each of these examples an assertion is made about the *relative frequency* with which a certain *property* occurs in a large *class* of objects or events. In the first example this is the property of yielding a one in the class of all throws of (fair) dice. In the second example the relative frequency of happy endings in the class of mystery films is observed; and in the last example a statement is made about the relative frequency with which the property of becoming an actor occurs in the class of actor's children.

The term "relative frequency" here used requires some closer examination. What exactly is the relative frequency of happy endings in mystery films? We may be tempted to give a very simple answer: of all the n mystery films ever produced, a certain number m does have a happy end. The relative frequency of happy endings then is m/n. However, this approach of defining the relative frequency of a property just as the fraction of elements that have the property must fail when the class of all objects or events under consideration is infinite. This might already be the case in the first example: that statement may very well be understood as making an assertion about all throws of a die, including those that will take place in the future. In this (potentially) infinite set we can not speak of the fraction of throws that result in a one. What is meant by the statement that the relative frequency of ones is $1/6$, is that in a large finite *random sample* of throws of a die the fraction of ones will be $1/6$. The canonical way of obtaining a random sample here being to take a die and throw it a large number of times.

But even if the class of all objects or events considered is finite, as in the second example, a stated relative frequency usually will have been obtained not by an examination of every individual object, but by drawing random samples from the domain. To say that mystery films end happily with a probability 0.7 may mean that we have systematically worked through a comprehensive film guide, and for each mystery film that we have found noted whether it has a happy or unhappy ending. It may also mean that we have watched a great number of mystery films on television and used this random sample as a basis for our assertion. The relative frequency of a specific property that we are going to observe then obviously depends on the way in which the random sample is obtained. Mystery films with a happy end will perhaps generally be more popular than those with tragic endings, and for that reason be more likely to be shown on TV. Thus, the relative frequency observed in an examination of a film guide and by watching television will differ to some extent.

We see that a statistical probability can not be taken to be an intrinsic property of the domain of objects or events to which it refers. It always, though often only implicitly, also refers to a specific *sampling method* according to which elements of the domain are observed.

While it is generally sufficient to imagine a statistical probability to refer to the relative frequency in a "large" sample drawn according to the given sampling method, this, of course, is not yet quite precise. To really obtain a working definition, we have to identify a statistical probability with the limit that the relative frequency is going to approach as the sample size tends towards infinity.

As yet, we always have been speaking of either objects or events to which probabilities refer, a class of objects or events in the case of statistical probabilities, and a single object or event in the case of subjective probabilities. However, it becomes rather tedious to always make explicit in our terminology the fact that probability statements can have such concrete subjects as the eggs used for making an omelette, or such abstract ones as the amount of precipitation on Sundays. Since we have seen that even a statement that, on the face of it, refers to a class of concrete objects (e.g. mystery films), in fact must be understood as referring to observations of these objects according to a specific rule, i.e. to events, in the sequel we adopt the term *event* to generally designate elements of the domain considered.

To summarize: a statistical probability for a certain property in a class of events is the limiting relative frequency with which this property is going to occur in large random samples of increasing size drawn from that class. It depends on the class of events under consideration as well as the specific sampling method used.

2.2 Syntax

We adopt the extension of first-order predicate logic for expressing statistical probabilities as given by Halpern[1990] and Bacchus[1990b]. The new basic construct of their language is a *statistical quantifier* $[\cdot]$ that allows us to construct from a given formula $\phi(v)$ a term

$$[\phi(v)]_v$$

representing the statistical probability that an event in the domain of discourse has property ϕ. More generally, we will not only be interested in the probabilities of properties of single events, but also in the frequency with which different events are related in a certain way, as for example in the statement "The probability that two actors working for the same studio have appeared in a film together is greater than 0.3". To provide for representations of statements of this sort, the statistical quantifier may also be used to quantify over more than one variable.

Furthermore, allowing for variables to remain free, we arrive at more general probability terms like

$$[\phi(u, v, w)]_{(u,v)}.$$

Allowing terms that represent probabilities in the language (which is not the only feasible way: Keisler's probability quantifiers [Keisler, 1985] represent a different approach) in addition to terms standing for objects of the actual domain of discourse, makes the resulting language two-sorted. Thus, we will use two distinct sets of variables $\{v_0, v_1, \dots w_0, w_1, \dots\}$ and $\{x_0, x_1, \dots y_0, y_1, \dots\}$, the first one ranging over elements of the domain, the second over the elements of an rc-field used to measure probabilities.

Apart from the usual logical connectives and quantifiers of first-order logic we also consider the vocabulary $\mathrm{S_{OF}}$ of ordered fields a fixed part of our language.

With these symbols and a (finite) vocabulary S containing symbols R,Q,... for relations on the domain, f,t,... for functions mapping domain elements to domain elements, and c,d,... for domain constants, we define domain-terms, field-terms, and formulas as in first-order logic to which is added one new syntax rule that allows to generate a field-term from a formula by means of the statistical quantifier:

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

- *Atomic domain formulas* are formulas of the form

  $$\mathsf{R}\,\mathrm{t}_1 \dots \mathrm{t}_k \quad \text{or} \quad \mathrm{t}_1 = \mathrm{t}_2,$$

  where 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 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 $\phi\langle \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{x}\rangle$ is a formula with free domain variables $\boldsymbol{v}$ and $\boldsymbol{w}$, and free field variables $\boldsymbol{x}$, then

  $$[\phi\langle \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{x}\rangle]_{\boldsymbol{w}}$$

  is a field-term in the variables $\boldsymbol{v}$ and $\boldsymbol{x}$.

- *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.

We denote by $\mathrm{FT}^{\sigma}_{\mathrm{S}}$ and $\mathrm{DT}^{\sigma}_{\mathrm{S}}$, respectively the sets of field-terms and domain-terms in the vocabulary S. $\mathrm{L}^{\sigma}_{\mathrm{S}}$ stands for the set of formulas in the vocabulary S.

Several abbreviations will be freely used to obtain more readable formulas. The usual conventions are used to eliminate superfluous parentheses. We write $\mathrm{t}_1 < \mathrm{t}_2$ as an abbreviation for $\mathrm{t}_1 \leq \mathrm{t}_2 \wedge \neg \mathrm{t}_2 \leq \mathrm{t}_1$, and $\mathrm{t}_1 = \mathrm{t}_2$ for $\mathrm{t}_1 \leq \mathrm{t}_2 \wedge \mathrm{t}_2 \leq \mathrm{t}_1$. $\mathrm{t}_1 \in [\mathrm{t}_2, \mathrm{t}_3]$ stands for $\mathrm{t}_1 \geq \mathrm{t}_2 \wedge \mathrm{t}_1 \leq \mathrm{t}_3$. Similarly for half-open and open intervals.

Any representation of a rational number (e.g. 0.55, 1/9,...) may be used as a constant symbol.

Notation for conditional probabilities is introduced as follows. Let t be a field-term containing the subterm

$$y \cdot [\phi(\boldsymbol{v},\boldsymbol{w},\boldsymbol{x}) \wedge \psi(\boldsymbol{v}', \boldsymbol{w}, \boldsymbol{x}')]_{\boldsymbol{w}}$$

(subterms of t are the terms from which t is constructed by multiplication and addition symbols; if t contains a subterm $[\psi(\ldots)]_{\ldots}$, and s is a term appearing in ψ, then this does not make s a subterm of t). The formula

$$\forall y(y \cdot [\psi(\boldsymbol{v}', \boldsymbol{w}, \boldsymbol{x}')]_{\boldsymbol{w}} = 1 \rightarrow \mathrm{t} \leq \mathrm{s})$$

then is abbreviated by

$$\mathrm{t}' \leq \mathrm{s}$$

where t' is t with the term $y \cdot [\phi(\boldsymbol{v},\boldsymbol{w},\boldsymbol{x}) \wedge \psi(\boldsymbol{v}', \boldsymbol{w}, \boldsymbol{x}')]_{\boldsymbol{w}}$ replaced by the expression

$$[\phi(\boldsymbol{v},\boldsymbol{w},\boldsymbol{x}) \mid \psi(\boldsymbol{v}', \boldsymbol{w}, \boldsymbol{x}')]_{\boldsymbol{w}}.$$

Thus, the intended meaning (not having defined semantics, we can not yet speak of the meaning proper) of the term $[\phi \mid \psi]$ is that of the quotient $[\phi \wedge \psi]/[\psi]$ when $[\psi] > 0$. In cases where $[\psi] = 0$, an atomic field formula containing the term $[\phi \mid \psi]$ expands to a formula that becomes vacuously true.

The examples of probabilistic statements mentioned so far in this section can now be formulated as L^{σ}-formulas in the following manner:

$$\begin{gathered}
[\mathsf{One}\, w \mid \mathsf{Die_throw}\, w]_w = 1/6 \\
[\mathsf{Happy_end}\, w \mid \mathsf{Mystery_film}\, w]_w = 0.7 \\
[\mathsf{Actor}\, w \mid \exists v(\mathsf{Actor}\, v \wedge \mathsf{father}(w) = v)]_w \geq 0.1 \\
[\exists v(\mathsf{Film}\, v \wedge \mathsf{In}\, uv \wedge \mathsf{In}\, wv) \mid \mathsf{Actor}\, u \wedge \mathsf{Actor}\, w \wedge \mathsf{studio}(u) = \mathsf{studio}(w)]_{(u,w)} > 0.3
\end{gathered}$$

2.2.1 Induction on $\mathrm{L}^{\sigma}_{\mathrm{S}}$

Sometimes we will want to prove assertions of the form "Every formula ϕ has property P". In first-order logic these proofs are often conducted by an induction on the structure of ϕ, i.e. it is first shown that property P holds when ϕ is atomic, and then it is proven that P holds when ϕ has been formed by a conjunction, negation, or quantification under the assumption that P holds for the subformulas of ϕ. The proof of the base case sometimes relies on an auxiliary statement of the form "Every term t has property P'.", which itself may have been proven by an induction on the structure of t.

An analogous proof method can be used for $\mathrm{L}^g_{\mathrm{S}}$. However, since here terms can, in turn, be constructed from formulas, the schema of the induction is a bit different. In order to prove a pair of assertions

(+) Every term $\mathrm{t} \in \mathrm{FT}^g_{\mathrm{S}}$ has property P′.
(++) Every formula $\phi \in \mathrm{L}^g_{\mathrm{S}}$ has property P.

we proceed as follows:

(a): Show that P′ holds for t and P holds for ϕ when t and ϕ are first-order, i.e. do not contain the statistical quantifier $[\cdot]$. This step may in turn be accomplished by standard inductions on the structures of t and ϕ.
(b): By an induction on the structure of t show that P′ holds for every field-term t. For the case $\mathrm{t} \equiv [\phi(\boldsymbol{v},\boldsymbol{w},\boldsymbol{x})]_{\boldsymbol{w}}$ assume that P holds for ϕ.
(c): By an induction on the structure of ϕ show that P holds for every formula ϕ. For the case $\phi \equiv \mathrm{t}_1 \leq \mathrm{t}_2$ assume that P′ holds for t_1 and t_2.

Steps (a)-(c) together provide a valid proof for (+) and (++): by (b) and (c) the validity of P′ and P for t and ϕ is reduced to the validity of the same properties for syntactically simpler terms or formulas. Specifically, the maximal depth to which nestings of the statistical quantifier occur in these terms and formulas is reduced by one. After finitely many recursion steps the problem is therefore reduced to first-order terms and formulas, which case is covered by (a).

2.3 Statistical Structures

The semantics for L^σ is defined in a similar way as in [Bacchus, 1990a]. The few points where the definitions given here vary from those supplied by Bacchus are listed at the end of this section.

The intuitive meaning of a term $[\phi(w)]_w$ is that of the probability that a randomly selected element of the domain (selected by the specific sampling method considered) has property ϕ, or, equivalently, belongs to the subset of the domain defined by ϕ. A term of the form $[\phi(\boldsymbol{w})]_{\boldsymbol{w}}$ with $\mid \boldsymbol{w} \mid = n > 1$ represents the probability that n randomly chosen elements of the domain are related by ϕ, or, equivalently, belong to the subset of n-tuples defined by ϕ. Thus, it is our intention to assign probability values to subsets of the domain and products of the domain. The semantics for the language $\mathrm{L}^g_{\mathrm{S}}$ will therefore be given by augmenting standard S-structures $\mathfrak{M}=(M,I)$ consisting of a domain M and an interpretation function I for the symbols in S, with an additional component for defining measures on subsets of (products of) the domain.

The general form of a structure $\mathfrak{M}$ for the interpretation of $\mathrm{L}^g_{\mathrm{S}}$ then is

$$\mathfrak{M} = (M, I, \mathfrak{F}, (\mathfrak{A}_n, \mu_n)_{n \in \mathbf{N}}), \tag{2.1}$$

with $\mathfrak{F}$ an rc-field and $(\mathfrak{A}_n, \mu_n)$ a measure algebra over M^n with probability values taken in $\mathfrak{F}$ $(n \in \mathbf{N})$.

Since for each n, the measure $\mu_n(A)$ is supposed to quantify the probability that a randomly chosen tuple of n elements belongs to the set $A \in \mathfrak{A}_n$, we will certainly have to demand that the sequence $(\mathfrak{A}_n, \mu_n)_n$ satisfies certain *consistency conditions* entailed by this interpretation of μ_n. Furthermore, a *closure condition* must be imposed on the $\mathfrak{A}_n$ that ensures that every subset of M^n definable in L_S^g belongs to $\mathfrak{A}_n$.

The closure condition can not be properly stated before the semantical relation between S-structures and formulas of L_S^g has been defined, because this relation specifies the set of definable subsets.

The consistency conditions, on the other hand, are purely measure theoretic conditions that are defined for $(\mathfrak{A}_n, \mu_n)_n$ without reference to L_S^g. All of the consistency conditions we now introduce are automatically satisfied if the sequence $(\mathfrak{A}_n, \mu_n)_n$ is in fact a sequence of product algebras and product measures, i.e. $\mathfrak{A}_n = \mathfrak{A}^n,\ \mu_n = \mu^n$ for $n \geq 1$.

The first condition states that μ_n is invariant under permutations.

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

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

Homogeneity states that it does not matter in which order we arrange the elements of our random sample: if, for instance, we consider two randomly chosen mystery films f_1 and f_2, then the probability that f_1 is a more famous film than f_2 is the same as the probability that f_2 is more famous than f_1.

The second condition concerns measurable rectangles.

Product property: For all $k, l \in \mathbf{N}$, $A \in \mathfrak{A}_k,\ B \in \mathfrak{A}_l$:

$$A \times B \in \mathfrak{A}_{k+l} \quad \text{and} \quad \mu_{k+l}(A \times B) = \mu_k(A)\mu_l(B).$$

The product property basically reflects the independence of the elements in the random sample: the probability that of two randomly selected mystery films the first one will have a happy end and the second one is, say, black and white is equal to the product of the probabilities that a single sample film will have a happy end, or be black and white.

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

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

Since the validity of the following property for products of σ-measures (precisely: σ-finite σ-measures) is the core of the proof of Fubini's theorem, we call it the Fubini property. (Fubini's theorem is a central theorem in measure- and integration theory about the interchangeability of the order of integration: $\int f(x,y)dxdy = \int f(x,y)dydx$.)

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

$$\sigma^I_{\boldsymbol{a}}(A) \in \mathfrak{A}_{n-k}, \tag{2.2}$$

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

$$A_{I,r} := \{\boldsymbol{a} \in M^I \mid \mu_{n-k}(\sigma^I_{\boldsymbol{a}}(A)) \geq r\} \in \mathfrak{A}_k, \tag{2.3}$$

and

$$\mu_n(A) \geq r\mu_k(A_{I,r}). \tag{2.4}$$

Figure 2.1 illustrates the Fubini property. The Fubini property, too, can be illustrated in the film example. Consider the property $\phi(\mathrm{f}_1, \mathrm{f}_2)$:"an actor of film f_1 has at some time in his or her career appeared in a film together with an actor of f_2". If we know that with a probability 0.1 a randomly selected film will have actors in it who have worked with such a large number of people so as to guarantee that in a second randomly selected film with a probability ≥ 0.2 one of these actors will appear, then the probability that a random pair of films has property ϕ must at least be $0.1 \cdot 0.2$.

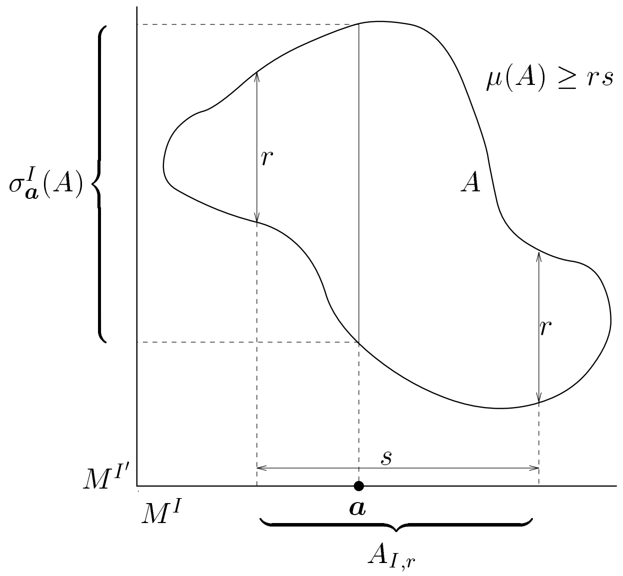

Figure 2.1: The Fubini property

It is easy to see that the Fubini property in fact makes the second provision of the product property redundant. Once it is assumed that $A \times B \in \mathfrak{A}_{k+l}$, the Fubini property can be used to derive that $\mu(A \times B) = \mu(A)\mu(B)$. The converse is not true: (2.4) can not be derived from the product property – even when the measurability conditions (2.2) and (2.3) are satisfied. An example where the product property holds, but (2.4) is violated will be given in section 2.4.4.

The Fubini property has been the last of the consistency conditions that we are going to impose on $(\mathfrak{A}_n, \mu_n)_n$. This begs the question: are these three conditions sufficient to guarantee that the measures (μ_n) behave according to their interpretation as representing the statistical probabilities governing the results of a sequence of n independent draws of events from the domain? It is hard to provide a conclusive affirmative answer to this question. In standard probability theory, product σ-measures are the canonical models for sequences of independent random draws. When we work only with algebras and (finitely additive) probability-measures, the sequence of product algebras and product measures $(\mathfrak{A}^n, \mu^n)$ is not quite rich enough for describing interesting outcomes of multiple draws, or, indeed, to satisfy the closure conditions below (if M is infinite, then the set $\{(a_1, a_2) \in M^2 \mid a_1 = a_2\}$ does not belong to $\mathfrak{A}^2$, for instance). We therefore have to use an augmented sequence $(\mathfrak{A}_n, \mu_n)_n \supseteq (\mathfrak{A}^n, \mu^n)_n$ in which additional sets become measurable. By demanding that the newly introduced sets $A \in \mathfrak{A}_n \setminus \mathfrak{A}^n$ and their measures satisfy the consistency conditions, it is ensured that they retain the central properties of a sequence of product (σ-) measures. This justifies some confidence that the sequence $(\mathfrak{A}_n, \mu_n)$ can not display any properties that are contrary to its intended meaning.

With the formulation of the consistency conditions, the semantical structures used to interpret $\mathrm{L}^g_{\mathrm{S}}$ are almost fully specified: they are going to be structures of the form 2.1 where $(\mathfrak{A}_n, \mu_n)_n$ satisfies the consistency conditions. As mentioned above, however, this is not yet quite sufficient: we will also have to demand that every subset of M^n that can be defined in $\mathrm{L}^g_{\mathrm{S}}$ belongs to $\mathfrak{A}_n$ – otherwise the interpretation of certain formulas in the language would be undefined in $\mathfrak{M}$.

The problem we face is that in order to avoid a certain circularity, we must not at this point simply declare that every definable set is measurable, because definability is explained in terms of the semantical $\models$ - relation, which has yet to be explained. The problem is circumvented by defining the satisfaction relation $\mathfrak{M} \models \phi$ between structures $\mathfrak{M}$ and formulas $\phi \in \mathrm{L}^g_{\mathrm{S}}$ for a wider class of structures than later will be used to define the semantics of ϕ.

Definition 2.3.1 Let S be a vocabulary, $\mathfrak{M} = (M, I, \mathfrak{F}, (\mathfrak{A}_n, \mu_n)_n)$ with (M, I) a standard model theoretic S-structure, $\mathfrak{F} = (\mathrm{F}, 0^{\mathrm{F}}, 1^{\mathrm{F}}, \leq^{\mathrm{F}}, \cdot^{\mathrm{F}}, +^{\mathrm{F}})$ a real closed field, and $(\mathfrak{A}_n, \mu_n)$ an $\mathfrak{F}$-measure algebra over M^n ($n \in \mathbf{N}$). Let γ be a variable assignment that maps domain-variables v into M and field-variables x into $\mathbf{F}$. We inductively define a mapping from the set of domain-terms into M, a partial mapping from the set of field-terms into $\mathbf{F}$, and a satisfaction relation $\models_\sigma$. We use the notation $\gamma[\boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r}]$ 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.

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 can not contain any field-terms as subterms.

Atomic domain formulas: Let $\phi(\boldsymbol{v}) \equiv \mathsf{R}\,\mathrm{t}_1(\boldsymbol{v}) \ldots \mathrm{t}_n(\boldsymbol{v})$ with $\mathsf{R} \in \mathrm{S}$ and domain-terms t_i. Then,

$$(\mathfrak{M}, \gamma) \models_\sigma \phi(\boldsymbol{v}) \;\text{ iff }\; ((\mathfrak{M}, \gamma)(\mathrm{t}_1) \ldots (\mathfrak{M}, \gamma)(\mathrm{t}_n)) \in I(\mathsf{R}).$$

Similarly for $\phi(\boldsymbol{v}) \equiv \mathrm{t}_1(\boldsymbol{v}) = \mathrm{t}_2(\boldsymbol{v})$ with domain terms $\mathrm{t}_1, \mathrm{t}_2$.

Boolean operators: Let $\phi(\boldsymbol{v}, \boldsymbol{x}) \equiv \psi(\boldsymbol{v}, \boldsymbol{x}) \wedge \chi(\boldsymbol{v}, \boldsymbol{x})$, then

$$(\mathfrak{M}, \gamma) \models_\sigma \phi(\boldsymbol{v}, \boldsymbol{x}) \;\text{ iff }\; (\mathfrak{M}, \gamma) \models_\sigma \psi(\boldsymbol{v}, \boldsymbol{x}) \text{ and } (\mathfrak{M}, \gamma) \models_\sigma \chi(\boldsymbol{v}, \boldsymbol{x}).$$

Similarly for $\phi(\boldsymbol{v},\boldsymbol{x}) \equiv \neg\psi(\boldsymbol{v},\boldsymbol{x})$.

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

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

For $\phi(\boldsymbol{v},\boldsymbol{x}) \equiv \exists y \psi(\boldsymbol{v},\boldsymbol{x}, y)$:

$$(\mathfrak{M},\gamma) \models_\sigma \phi(\boldsymbol{v},\boldsymbol{x}) \text{ iff } \exists r \in \mathbf{F} \quad (\mathfrak{M},\gamma[y/r]) \models_\sigma \psi(\boldsymbol{v},\boldsymbol{x}, y).$$

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

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

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

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

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

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

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

Atomic field formulas: Let $\phi(\boldsymbol{v},\boldsymbol{x}) \equiv \mathrm{t}_1(\boldsymbol{v},\boldsymbol{x}) \leq \mathrm{t}_2(\boldsymbol{v},\boldsymbol{x})$. Then $(\mathfrak{M},\gamma) \models_\sigma \phi(\boldsymbol{v},\boldsymbol{x})$ iff $(\mathfrak{M},\gamma)(\mathrm{t}_1)$ and $(\mathfrak{M},\gamma)(\mathrm{t}_2)$ are defined, and $(\mathfrak{M},\gamma)(\mathrm{t}_1) \leq^{\mathrm{F}} (\mathfrak{M},\gamma)(\mathrm{t}_2)$.

In this definition, the validity of $(\mathfrak{M},\gamma) \models_\sigma \phi(\boldsymbol{v},\boldsymbol{x})$ only depends on the values of γ for variables belonging to $\boldsymbol{v}$ or $\boldsymbol{x}$. For this reason, when $\gamma(\boldsymbol{v}) = \boldsymbol{a}$ and $\gamma(\boldsymbol{x}) = \boldsymbol{r}$, we may also write $(\mathfrak{M},\boldsymbol{v}/\boldsymbol{a},\boldsymbol{x}/\boldsymbol{r}) \models_\sigma \phi(\boldsymbol{v},\boldsymbol{x})$ for $(\mathfrak{M},\gamma) \models_\sigma \phi(\boldsymbol{v},\boldsymbol{x})$.

The interpretation of certain field-terms is undefined in $\mathfrak{M}$ when there exist definable subsets of M^n that are not measurable. The next definition introduces a more compact notation for subsets defined by a formula $\phi \in \mathrm{L}_{\mathrm{S}}^{g}$ in a structure $\mathfrak{M}$.

Definition 2.3.2 Let $\mathfrak{M}$ be as in definition 2.3.1, $\phi\langle\boldsymbol{v},\boldsymbol{w},\boldsymbol{x},\boldsymbol{y}\rangle \in \mathrm{L}_{\mathrm{S}}^{g}$, $\boldsymbol{a} \in M^{|\boldsymbol{v}|}$, $\boldsymbol{r} \in \mathbf{F}^{|\boldsymbol{x}|}$. The subset of $M^{|\boldsymbol{w}|} \times \mathbf{F}^{|\boldsymbol{y}|}$ defined in $\mathfrak{M}$ by ϕ with parameters $\boldsymbol{a}$,$\boldsymbol{r}$ then is

$$(\mathfrak{M},\boldsymbol{v}/\boldsymbol{a},\boldsymbol{x}/\boldsymbol{r})(\phi\langle\boldsymbol{v},\boldsymbol{w},\boldsymbol{x},\boldsymbol{y}\rangle) := \{(\boldsymbol{b},\boldsymbol{s}) \mid (\mathfrak{M},\boldsymbol{v}/\boldsymbol{a},\boldsymbol{w}/\boldsymbol{b},\boldsymbol{x}/\boldsymbol{r},\boldsymbol{y}/\boldsymbol{s}) \models_\sigma \phi\langle\boldsymbol{v},\boldsymbol{w},\boldsymbol{x},\boldsymbol{y}\rangle\}. \tag{2.5}$$

Note that the set defined by $\phi\langle\boldsymbol{v},\boldsymbol{w},\boldsymbol{x},\boldsymbol{y}\rangle$ depends on the order implicitly given to the variables of ϕ by the notation $\phi\langle\boldsymbol{v},\boldsymbol{w},\boldsymbol{x},\boldsymbol{y}\rangle$. Writing $\phi\langle\boldsymbol{v},\pi\boldsymbol{w},\boldsymbol{x},\pi'\boldsymbol{y}\rangle$, with π (π') a permutation of $\{1,\ldots,\mid \boldsymbol{w} \mid\}$ $(\{1,\ldots,\mid \boldsymbol{y} \mid\})$, results in a set $(\mathfrak{M},\boldsymbol{v}/\boldsymbol{a},\boldsymbol{x}/\boldsymbol{r})(\phi\langle\boldsymbol{v},\pi\boldsymbol{w},\boldsymbol{x},\pi'\boldsymbol{y}\rangle)$ that is the permutation $\pi(A) \times \pi'(R)$ of $A \times R = (\mathfrak{M},\boldsymbol{v}/\boldsymbol{a},\boldsymbol{x}/\boldsymbol{r})(\phi\langle\boldsymbol{v},\boldsymbol{w},\boldsymbol{x},\boldsymbol{y}\rangle)$.

Example 2.3.3 Let R be a binary relation symbol, $\phi \equiv \mathsf{R}\,uv$. Then

$$\mathfrak{M}(\phi\langle u,v\rangle) = I(\mathsf{R}) \subseteq M^2,$$

while

$$\mathfrak{M}(\phi\langle v,u\rangle) = \{(a,b) \mid (b,a) \in I(\mathsf{R})\} \subseteq M^2.$$

Definition 2.3.2 has been stated in such a way (by writing $\phi\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x},\boldsymbol{y}\rangle$, not $\phi(\boldsymbol{v},\boldsymbol{w},\boldsymbol{x},\boldsymbol{y})$) that a formula with exactly k free domain variables that are not used as parameters, and l free field variables not used as parameters, always defines a subset of $M^k \times \mathbf{F}^l$. However, by (2.5) just as well a subset of $M^{|\boldsymbol{w}|} \times \mathbf{F}^{|\boldsymbol{y}|}$ is defined when it is not presumed that ϕ actually contains all the variables in $\boldsymbol{w}$ and $\boldsymbol{y}$. Hence, for any given $\phi(\boldsymbol{v},\boldsymbol{w},\boldsymbol{x},\boldsymbol{y})$ that may or may not actually contain all the variables in $\boldsymbol{w}$ and $\boldsymbol{y}$, we may also loosely speak of the set

$$(\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\phi(\boldsymbol{v},\boldsymbol{w},\boldsymbol{x},\boldsymbol{y})) \subseteq M^{|\boldsymbol{w}|} \times \mathbf{F}^{|\boldsymbol{y}|}$$

defined by $\phi(\boldsymbol{v},\boldsymbol{w},\boldsymbol{x},\boldsymbol{y})$, which, strictly according to definition 2.3.2, is the set defined by

$$\hat{\phi}\langle \boldsymbol{v}', \boldsymbol{w}, \boldsymbol{x}', \boldsymbol{y}\rangle :\equiv \phi(\boldsymbol{v},\boldsymbol{w},\boldsymbol{x},\boldsymbol{y}) \wedge \bigwedge_{w \in \boldsymbol{w}\setminus\boldsymbol{w}'} w = w \wedge \bigwedge_{y \in \boldsymbol{y}\setminus\boldsymbol{y}'} y = y,$$

where $\boldsymbol{v}' \subseteq \boldsymbol{v}$, $\boldsymbol{w}' \subseteq \boldsymbol{w}$, $\boldsymbol{x}' \subseteq \boldsymbol{x}$, and $\boldsymbol{y}' \subseteq \boldsymbol{y}$ are the variables that actually appear in ϕ.

For $\mathfrak{M}$ as in definition 2.3.1 it is clearly equivalent that $(\mathfrak{M},\gamma)(\mathrm{t})$ is defined for every variable assignment γ and every $\mathrm{t} \in \mathrm{FT}^g_{\mathrm{S}}$, and that every subset of M^n definable with parameters from M and $\mathbf{F}$ by some $\phi \in \mathrm{L}^g_{\mathrm{S}}$ belongs to $\mathfrak{A}_n$. Any of these two statements would make a suitable formulation of the closure condition that, in addition to the consistency conditions, is required to make $\mathfrak{M}$ a structure for the interpretation of $\mathrm{L}^g_{\mathrm{S}}$. The following definition makes use of the second one.

Definition 2.3.4 Let $\mathfrak{M}$=$(M, I, \mathfrak{F}, (\mathfrak{A}_n, \mu_n)_n)$ as in definition 2.3.1. $\mathfrak{M}$ is a *statistical* S-*structure* iff

- $(\mathfrak{A}_n, \mu_n)_n$ satisfies homogeneity, the product-, and the Fubini-property,
- For each n, $A \subseteq M^n$: if A is definable with parameters in $\mathrm{L}^g_{\mathrm{S}}$, then $A \in \mathfrak{A}_n$.

If $\mathfrak{F} = \mathfrak{R}$, then $\mathfrak{M}$ is called a *real-valued* statistical structure.

Example 2.3.5 Let $\mathfrak{M}$ be a statistical S-structure, $\boldsymbol{v} = (v_1, \ldots, v_n)$, $\boldsymbol{w} = (v_1, \ldots, w_n)$. Let

$$\phi\langle \boldsymbol{v},\boldsymbol{w}\rangle :\equiv \bigwedge_{i=1}^{n} v_i = w_i.$$

Let $\boldsymbol{a} \in M^n$. Then

$$(\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a})(\phi\langle \boldsymbol{v},\boldsymbol{w}\rangle) = \{\boldsymbol{a}\} \subseteq M^n.$$

Hence, in a statistical S-structure, $\mathfrak{A}_n$ contains all singleton sets, and, consequently, all finite and co-finite sets.

Example 2.3.6 Let $\mathfrak{M}$ be a statistical S-structure, $\boldsymbol{v} = (v_1, \ldots, v_n)$, $j, k \in \{1, \ldots, n\}$, $j \neq k$. Let

$$\phi^H(\boldsymbol{v}) :\equiv v_j = v_k.$$

Then $\mathfrak{M}(\phi^H(\boldsymbol{v}))$ is the "hyperplane"

$$H = \{\boldsymbol{a} \in M^n \mid \boldsymbol{a}(j) = \boldsymbol{a}(k)\},$$

which therefore must belong to $\mathfrak{A}_n$.

Example 2.3.7 Let (M, I) be a standard S-structure, $a_1, a_2, \ldots \in M$, $p_1, p_2, \ldots \in \mathbf{R}$ with $p_i \geq 0$ and $\sum p_i = 1$. For each $n \geq 1$ a real-discrete measure (cf. example 1.2.7) is defined on $\mathfrak{A}_n := 2^{(M^n)}$ via

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

It is easy to see that $(\mathfrak{A}_n, \mu_n)_n$ satisfies the consistency conditions, so that $(M, I, \mathfrak{F}, (\mathfrak{A}_n, \mu_n)_n)$ is a statistical S-structure. We refer to structures of this form as *real-discrete structures*.

Example 2.3.8 Let $\mathfrak{M}$ be a statistical S-structure with

$$\mathfrak{M} \models_\sigma \forall v[v = w]_w = 0 \quad \equiv: \phi^{\text{cont}}.$$

Then $\mu_1(\{a\}) = 0$ for all $a \in M$, and, by the product property, also $\mu_n(\{\boldsymbol{a}\}) = 0$ for all $\boldsymbol{a} \in M^n$. In other words, μ_n is a continuous measure for all n (cf. example 1.2.8). Models of ϕ^{cont} therefore are called *continuous structures*.

From definitions 2.3.1 and 2.3.4 in the usual way we arrive at the definitions for the *semantic entailment* relation between L^σ-formulas, which is defined in two versions.

Definition 2.3.9 Let $\Phi \subseteq \mathrm{L}^\sigma_{\mathrm{S}}$, $\phi \in \mathrm{L}^\sigma_{\mathrm{S}}$. ϕ is σ-entailed by Φ, written $\Phi \models_\sigma \phi$, if every statistical S-structure that is a model of Φ, also is a model of ϕ. Also, we use the notation $\Phi \models^{\mathbf{R}}_\sigma \phi$ if every real-valued model of Φ is a model of ϕ.

Definition 2.3.10 We write $\mathscr{L}^\sigma$ for the logic defined by the language L^σ and the entailment relation $\models_\sigma$.

Example 2.3.11 Let $\mathfrak{M}$ be a continuous statistical structure. Let ϕ^H and H as in example 2.3.6. Let $I = \{1, \ldots, n\} \setminus \{k\}$, $\boldsymbol{a} \in M^I$. Then

$$\sigma^I_{\boldsymbol{a}}(H^c) = M \setminus \{\boldsymbol{a}(j)\} \in \mathfrak{A}_1.$$

By the continuity of μ_1:

$$\mu_1(\sigma^I_{\boldsymbol{a}}(H^c)) = \mu_1(M) - \mu_1(\{\boldsymbol{a}(j)\}) = 1.$$

Since $\boldsymbol{a}$ was arbitrary, we obtain

$$(H^c)_{I,1} = M^I$$

(recall how $(H^c)_{I,1}$ is defined by (2.3)), and by the Fubini-property

$$\mu_n(H^c) = 1.$$

Thus, the continuity of μ_1 entails that "hyperplanes" have zero probability:

$$\phi^{\mathrm{cont}} \models_\sigma [\phi^H(\boldsymbol{v})]_{\boldsymbol{v}} = 0. \tag{2.6}$$

Example 2.3.12 Let R,S be unary relation symbols, $r \in \mathbf{Q}$. Then

$$[\mathsf{R}w]_w = 0 \models_\sigma [\mathsf{S}w \mid \mathsf{R}w]_w < r \wedge [\mathsf{S}w \mid \mathsf{R}w]_w > r.$$

This is immediate from the definition of the conditional probability expression. $[\mathsf{S}w \mid \mathsf{R}w]_w < r$, for example, being an abbreviation for

$$\forall y(y \cdot [\mathsf{R}w]_w = 1 \rightarrow y \cdot [\mathsf{S}w \wedge \mathsf{R}w]_w < r),$$

is entailed by $[\mathsf{R}w]_w = 0$ for any r because the antecedent $y \cdot [\mathsf{R}w]_w = 1$ is not satisfiable for any y.

The same is true for any more complicated field-term t containing a conditional probability expression $[\phi \mid \psi]$: from the premise $[\psi] = 0$ any inequality $\mathrm{t} \leq \mathrm{s}$ can be inferred. Thus, when the conditioning set ψ has probability zero, the conditional probability $[\phi \mid \psi]$ behaves like an undefined term that makes any inequality valid in which it appears. This does not lead to any unwanted contradictions because the expression $[\phi \mid \psi]$ does not directly abbreviate a field-term, so that from $[\phi \mid \psi] < r \wedge [\phi \mid \psi] > r$ we can not infer $\exists y(y < r \wedge y > r)$.

The following lemma states that the semantics for L^σ is an extension of the semantics of standard first-order logic.

Lemma 2.3.13 Let Φ, ϕ be first-order. Then

$$\Phi \models_\sigma \phi \quad \text{iff} \quad \Phi \models \phi.$$

Proof: For the right to left direction it is sufficient to observe that for first-order formulas ϕ only the standard part (M, I) of a statistical S-structure is used for defining the relation $(\mathfrak{M}, \gamma) \models_\sigma \phi$ in the same manner as in first-order logic. For the converse direction, it must be noted that every standard structure (M, I) can be extended to a real-discrete statistical S-structure $\mathfrak{M}$ (just assign probability 1 to an arbitrary element of M), and that for such $\mathfrak{M}$ and first-order formulas ϕ

$$(\mathfrak{M}, \gamma) \models_\sigma \phi \;\text{ iff }\; ((M, I), \gamma) \models \phi.$$

$\square$

Within the present work we are essentially concerned with the standard notion of probability as a real number. Defining semantics for L^σ that allows for probabilities taken in arbitrary rc-fields therefore must be seen as an approximation only of the intended meaning of L^σ-sentences, and the entailment relation $\models_\sigma$ only is an approximation of the relation $\models_\sigma^{\mathbf{R}}$ which characterizes the inferences we would really like to draw from L^σ-formulas. This approximation is correct in the sense that $\Phi \models_\sigma \phi$ implies $\Phi \models_\sigma^{\mathbf{R}} \phi$, but not complete because the converse does not hold. The reason for nonetheless working with the weaker semantics based on rc-fields is that for this case we can find a complete proof-system (see [Bacchus, 1990a] and section 2.5 below). For the entailment relation $\models_\sigma^{\mathbf{R}}$, on the other hand, Abadi and Halpern [1989] have shown that no complete proof system exists.

Besides obtaining completeness results, there exist other reasons that can motivate the study of probabilities other than expressed by real numbers. Particularly, as a model for subjective probabilities, simpler structures than the reals can support the notion of a qualitative probability (e.g. [Aleliunas, 1990]). Probabilities in elementary extensions of $\mathfrak{R}$, on the other hand, might be helpful for developing a uniform theory of probabilistic and default reasoning ([Weydert, 1995]).

Giving up real-valued probabilities makes it also necessary to give up σ-additivity. This is not too big a sacrifice, however, because in the context of the finitary language L^σ, where definability is preserved under finite, but not under countable unions, finite additivity seems to be the more natural concept anyway.

The only concrete examples of statistical S-structures that we have met so far are the real-discrete structures of example 2.3.7. For continuous structures (example 2.3.8), on the other hand, no explicit description has been given so far, so that at this point it is not even clear that such structures exist.

A very interesting and natural type of continuous structures for reasoning about geometric objects might be given by taking an interval in $\mathbf{R}$, say [0,1], as the domain. On n-dimensional products of this interval the σ-algebra $\bar{\mathfrak{B}}[0,1]^n$ of Lebesgue measurable sets with the Lebesgue measure $\bar{\lambda}^n$ is defined. (The σ-algebra of Lebesgue measurable sets is the completion of the σ-algebra of Borel sets, i.e. the σ-algebra generated by the Borel sets and all subsets of Borel sets with Lebesgue measure 0. The Borel sets, in turn, are the elements of the σ-algebra generated by all n-dimensional intervals. The Lebesgue measure is the measure that assigns to each n-dimensional interval its volume.) Letting $(\mathfrak{A}_n, \mu_n) = (\bar{\mathfrak{B}}[0,1]^n, \bar{\lambda}^n)$ would then yield a suitable structure in which we can describe geometric objects of various dimensions, whose "probability" is given by their "volume".

Unfortunately, we encounter very deep problems when we try to construct structures in this way. The system of measurable sets not being closed under projections (as is very easy to see: for any nonmeasurable $A \subset \mathbf{R}$, the set $\{0\} \times A \subset \mathbf{R}^2$ is measurable with projection A on the second component), the closure condition would be violated if we attempted to interpret some relation symbol by a measurable set with a nonmeasurable projection, because the projection of a definable set is definable via quantification.

But what if we interpret the symbols in S by sufficiently simple measurable sets, specifically Borel sets (for function symbols f this means: the graph of f is a Borel set)? Even for pure first-order logic it is unknown whether, starting from Borel sets, it is possible to define nonmeasurable sets. This to be the case, however, is considered to be extremely unlikely, so that *projective determinacy* (PD), an axiom for set theory that implies the measurability of definable subsets of $\mathbf{R}^n$, is widely considered an acceptable extension of ZFC (see [Martin, 1977] for a survey).

In spite of this being a fascinating subject, it is impossible at this place to explore the situation that is created by considering definability in L^σ rather than first-order logic, specifically with regard to the question whether PD will still guarantee measurability for sets definable in L^σ starting from Borel sets.

There is a special case, however, for which it is very easy to construct continuous statistical S-structures over a domain M equipped with a measure algebra $(\mathfrak{A}, \mu)$ with continuous μ: when S is *monadic*, i.e. only contains constant- and unary relation symbols, then the class of definable subsets in M^n only contains sets of a very simple form that can be assigned a measure in a canonical way. The next section gives the details of this construction.

To conclude this section, some differences between the material presented here and that contained in [Bacchus, 1990a] should be pointed out: first, unlike Bacchus, we make no restrictions to structures with countable domains. On the other hand, rather than allowing probabilities to take values in any totally ordered field, we only consider real closed fields, which provide the best first-order definable approximation of the real numbers. Bacchus explicitly demands homogeneity and the product property for his semantical structures, but not the Fubini property. This seems to be an involuntary rather than an intentional omission: in his proof theory he uses an axiom that is validated only by the Fubini- but not by the product property. Bacchus introduces the conditional probability expression as a field-term that is defined to be zero, when the probability of the conditioning set is zero. Finally, in order to slightly simplify our exposition, we here have made no provisions for "measuring functions" – functions that take elements of the domain as arguments and return field-values.

2.4 Structures for Monadic Languages

In this section a constructive description of a class of continuous S-structures for monadic vocabularies S is given. It is shown that for such vocabularies a "1-dimensional" structure $(M, I, \mathfrak{F}, (\mathfrak{A}, \mu))$, with $(\mathfrak{A}, \mu)$ a continuous measure algebra over M that contains the interpretations $I(\mathsf{R})$ of the relation symbols in S, always can be extended to a statistical S-structure $(M, I, \mathfrak{F}, (\mathfrak{A}^{\mathrm{m}}_n, \mu^{\mathrm{m}}_n)_n)$ with $(\mathfrak{A}^{\mathrm{m}}_1, \mu^{\mathrm{m}}_1) = (\mathfrak{A}, \mu)$. Particularly, it will be the case that for $(\mathfrak{A}, \mu) = (\mathfrak{B}, \lambda)$ the σ-algebra of Borel-sets, each $(\mathfrak{A}^{\mathrm{m}}_n, \mu^{\mathrm{m}}_n)$ is a subalgebra of $(\mathfrak{B}^n, \lambda^n)$.

The main reason why we here will spend some effort on the construction of these statistical structures is the necessity of showing that there exist other types of statistical S-structures than the simple real-discrete structures. If these latter type of structures were the only ones that existed, then the whole complicated definition of statistical structures in terms of the consistency and closure conditions could be dispensed with, and we could define a statistical structure as real-discrete structure right away.

Intuitively it should be very easy to find suitable $(\mathfrak{A}^{\mathrm{m}}_n, \mu^{\mathrm{m}}_n)$ for building our statistical structures: sets definable in the monadic language will basically have to be boolean combinations of measurable rectangles $A_1 \times \ldots \times A_n$ $(A_i \in \mathfrak{A})$, i.e. belong to the product algebra $\mathfrak{A}^n$. It turns out, that this is essentially true, only that the situation is somewhat complicated by the necessity to also accommodate subsets definable by the equality predicate in $\mathfrak{A}_n^{\mathrm{m}}$, so that letting $\mathfrak{A}_n^{\mathrm{m}} := \mathfrak{A}^n$ is not quite sufficient.

While the structures defined in this section really are of a very simple nature, the proofs that they have the properties we need are somewhat tedious, and the reader may wish to skip them at a first reading.

2.4.1 Defining $\mathfrak{A}_n^{\mathrm{m}}$ and μ_n^{m}

Apart from the product algebra $\mathfrak{A}^n$, each $\mathfrak{A}_n^{\mathrm{m}}$ must contain the subsets definable via the equality predicate. The most basic sets of this kind are of the form

$$E_{i,j} := \{\boldsymbol{a} \in M^n \mid \boldsymbol{a}(i) = \boldsymbol{a}(j)\} \quad i, j \in \{1, \dots, n\}.$$

The algebra $\mathfrak{X}_n$ generated by the system $\{E_{i,j} \mid i, j \in \{1, \dots, n\}\}$ then must also be contained in $\mathfrak{A}_n^{\mathrm{m}}$. $\mathfrak{X}_n$ is finite. The atoms of $\mathfrak{X}_n$ are the nonempty sets of the form

$$\bigcap_{i,j} \tilde{E}_{i,j}, \quad \tilde{E}_{i,j} \in \{E_{i,j}, (E_{i,j})^c\}.$$

It is easy to see that an intersection of this form is nonempty, iff

$$Jij \;:\Leftrightarrow\; \tilde{E}_{i,j} = E_{i,j}$$

is an equivalence relation, so that the atoms of $\mathfrak{X}_n$ are the sets

$$X(J) := \{\boldsymbol{a} \in M^n \mid \boldsymbol{a}(i) = \boldsymbol{a}(j) \Leftrightarrow (i, j) \in J\}$$

defined by equivalence relations J on $\{1, \dots, n\}$. By theorem 1.2.13 we know that the algebra

$$\mathfrak{A}_n^{\mathrm{m}} := \mathfrak{A}(\mathfrak{A}^n, \mathfrak{X}_n) \tag{2.7}$$

is disjointly generated by the system

$$\tilde{\mathfrak{E}}_n := \{A \cap X(J) \mid A = \times_{i=1}^n A_i,\ A_i \in \mathfrak{A},\ J \text{ an equivalence relation on } \{1, \dots, n\}\}.$$

Example 2.4.1 On $\{1, 2\}$ there are only two equivalence relations: J_1 defined by $(1, 2) \in J_1$, and J_2 defined by $(1, 2) \notin J_2$. Figure 2.2 shows two sets $(A_1 \times A_2) \cap X(J_1)$ and $(B_1 \times B_2) \cap X(J_2)$ belonging to $\tilde{\mathfrak{E}}_2$.

The generating system $\tilde{\mathfrak{E}}_n$ provides the simplest and most intuitive description of the algebra $\mathfrak{A}_n^{\mathrm{m}}$. To facilitate some of the proofs to come, it is advisable, however, to work with generating systems $\mathfrak{E}_n \subset \tilde{\mathfrak{E}}_n$ that only contain sets with some specific properties. To define these properties, and at the same time to show that the subsystem of $\tilde{\mathfrak{E}}_n$ containing the sets with these properties still generates $\mathfrak{A}_n^{\mathrm{m}}$, we take an arbitrary element $E = \times A_i \cap X(J) \in \tilde{\mathfrak{E}}_n$, $E \neq \emptyset$, and show how it can be divided into a finite collection of disjoint $E_j \in \tilde{\mathfrak{E}}_n$ of a simpler structure.

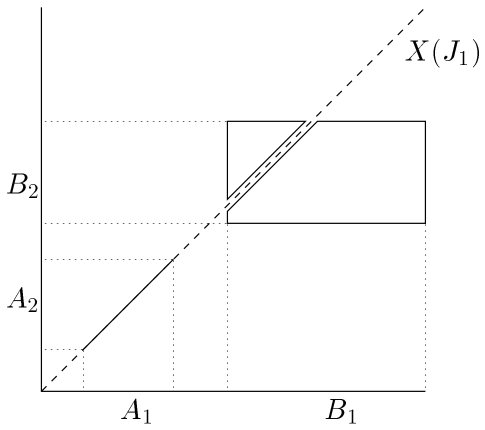

Figure 2.2: Generating sets of $\mathfrak{A}_2^{\mathrm{m}}$

First, for each $i \in 1, \dots, n$ let

$$A'_i := \bigcap_{\{j \mid (i,j) \in J\}} A_j$$

Then clearly $\times_{i=1}^{n} A'_i \cap X(J)$ is an alternative representation of E for which

$$(i,j) \in J \Rightarrow A'_i = A'_j. \tag{2.8}$$

Now let $I^f(E)$ be the set of indices $i \in \{1, \dots, n\}$ with finite A'_i (the notation $I^f(E)$ is justified, because $I^f(E)$ is indeed independent of the representation chosen for E, provided this representation satisfies (2.8)). For each $\boldsymbol{a} \in \times_{i \in I^f(E)} A'_i$ let

$$E(\boldsymbol{a}) = \times_{i=1}^{n} A''_i \cap X(J),$$

where

$$A''_i = \begin{cases} \{\boldsymbol{a}(i)\} & \text{if } i \in I^f(E) \\ A'_i \setminus \{\boldsymbol{a}(j) \mid j \in I^f(E)\} & \text{else.} \end{cases}$$

Then

$$E = \dot{\bigcup_{\boldsymbol{a} \in \times_{i \in I^f(E)} A'_i}} E(\boldsymbol{a}).$$

Each nonempty $E(\boldsymbol{a})$ is an element of $\tilde{\mathfrak{E}}_n$ of the form $\times A''_i \cap X(J)$ such that (2.8) holds (this property is preserved when A'_i is replaced by A''_i !), and furthermore

$$\forall i \quad |A''_i| = 1 \text{ or } |A''_i| = \infty \tag{2.9}$$

$$\forall i,j \quad |A''_i| = \infty \text{ and } |A''_j| = 1 \;\Rightarrow\; A''_j \not\subset A''_i \tag{2.10}$$

Calling a set $E \in \tilde{\mathfrak{E}}_n$ *simple* if it admits a representation by A_i's that satisfy (2.8)-(2.10), we obtain that

$$\mathfrak{E}_n := \{E \in \tilde{\mathfrak{E}}_n \mid E \text{ simple}\} \tag{2.11}$$

disjointly generates $\mathfrak{A}_n^{\mathrm{m}}$.

Example 2.4.2 Let $M = \mathbf{N}$, $\mathfrak{A} = 2^{\mathbf{N}}$. Define

$$\begin{aligned} A_1 &= \{n \mid n \text{ divisible by } 3\} \\ A_2 &= \{n \mid n \text{ divisible by } 5\} \\ A_3 &= \{15\} \\ A_4 &= \{17, 19\} \\ A_5 &= \{n \mid n \text{ prime}\} \end{aligned}$$

Let J be given by the equivalence classes $\{1,2\}, \{3\}$, and $\{4,5\}$. Let $E = \times_{i=1}^5 A_i \cap X(J)$. Taking the intersections of the A_i's belonging to the same equivalence class, we obtain

$$\begin{aligned} A_1' = A_2' &= \{n \mid n \text{ divisible by } 15\} \\ A_3' &= \{15\} \\ A_4' = A_5' &= \{17, 19\} \end{aligned}$$

and $I^f(E) = \{3, 4, 5\}$. The only elements $\boldsymbol{a} \in \times_{i \in I^f(E)} A_i'$ for which $E(\boldsymbol{a})$ is nonempty, are $\boldsymbol{a}$=(15,17,17) and $\boldsymbol{a}' = (15, 19, 19)$ with

$$\begin{aligned} E(\boldsymbol{a}) &= [A_1' \setminus \{15\} \times A_2' \setminus \{15\} \times \{15\} \times \{17\} \times \{17\}] \cap X(J), \\ E(\boldsymbol{a}') &= [A_1' \setminus \{15\} \times A_2' \setminus \{15\} \times \{15\} \times \{19\} \times \{19\}] \cap X(J), \end{aligned}$$

so that $E = E(\boldsymbol{a}) \dot\cup E(\boldsymbol{a}')$.

Lemma 2.4.3 For $E \in \mathfrak{E}_n \setminus \{\emptyset\}$ there exists only one representation $\times A_i \cap X(J)$ that satisfies (2.8)-(2.10).

Proof: Let $\times B_i \cap X(J')$ be an alternative representation of E. Clearly, then $J' = J$ and $B_i = A_i$ for all A_i with $\mid A_i \mid = 1$. From (2.8) and (2.10) it follows easily that for every element a from an infinite A_i there exists $\boldsymbol{a} \in \times A_i \cap X(J)$ such that $\boldsymbol{a}(i) = a$, so that $a \in B_i$ follows. In the same manner we obtain $B_i \subseteq A_i$, proving the lemma. □

In the sequel, when we consider sets $E = \times A_i \cap X(J) \in \mathfrak{E}_n$, it will always be assumed that the given representation of E is the one that satisfies (2.8)-(2.10).

According to theorem 1.2.10, to define a probability measure μ_n^{m} on $\mathfrak{A}_n^{\mathrm{m}}$, we only have to define an $\mathbf{F}^+$- valued additive function μ_n^0 on $\mathfrak{E}_n$ such that $\mu_n^0(M^n) := \sum_J \mu_n^0(M^n \cap X(J)) = 1$. In view of example 2.3.11, and the fact that the resulting measures are supposed to have the product property, it turns out that there can only be one way to define μ_n^0.

Generalizing the definition of $X(J_2)$ in example 2.4.1, we let $J_n^{\neq}$ be the minimal equivalence relation on $\{1,\dots,n\}$, i.e.

$$J_n^{\neq} = \{(i,i) \mid i = 1,\dots,n\},$$

so that

$$X(J_n^{\neq}) = \{\boldsymbol{a} \in M^n \mid \boldsymbol{a}(i) \neq \boldsymbol{a}(j)\ \forall i \neq j\}.$$

Define on $\mathfrak{E}_n \setminus \{\emptyset\}$

$$\mu_n^0(A \cap X(J)) = \begin{cases} \mu^n(A) & \text{if } J = J_n^{\neq} \\ 0 & \text{else} \end{cases} \tag{2.12}$$

with $\mu^n(A) = \mu^n(\times_{i=1}^n A_i) = \prod_{i=1}^n \mu(A_i)$ the product measure on $\mathfrak{A}^n$, and let $\mu_n^0(\emptyset) := 0$.

To improve readability, in the sequel the subscript n will be dropped from $J_n^{\neq}$, μ_n^0, $\mathfrak{A}_n^{\mathrm{m}}, \dots$ when in the given context an explicit reference to the dimension is unnecessary.

Lemma 2.4.4 μ_n^0 is additive.

Proof: Let $E = A \cap X(J),\ E_1 = B \cap X(J'),\ E_2 = C \cap X(J'')\ \in \mathfrak{E}_n \setminus \emptyset$ with $E = E_1 \dot{\cup} E_2$.

It immediately follows that $J = J' = J''$, and that $\mu_n^0(E) = \mu_n^0(E_1) + \mu_n^0(E_2)$ holds if $J \neq J^{\neq}$.

Suppose, then, that $J = J^{\neq}$. If any of the factors A_i of A is finite, then so are the corresponding factors B_i and C_i, and, by the continuity of μ, we have $\mu^0(E) = \mu^0(E_1) = \mu^0(E_2) = 0$.

The case in which all the A_i are infinite requires a little work. We are going to show that then there exists an $i_0 \in \{1,\dots,n\}$ such that

(i) $\forall i \neq i_0 \quad C_i \subseteq B_i$

(ii) $B_{i_0} \cap C_{i_0} = \emptyset$

(iii) $\forall i \neq i_0 \quad B_i \setminus C_i$ is finite

(iv) $\forall i \neq i_0 \quad C_i$ is infinite

(modulo interchanging the roles of B and C throughout (i)-(iv)). Using (i)-(iv) it is then shown that

$$\times_{i \neq i_0} C_i \times (B_{i_0} \dot{\cup} C_{i_0}) \ \subseteq A \ \subseteq \times_{i \neq i_0} B_i \times (B_{i_0} \dot{\cup} C_{i_0}). \tag{2.13}$$

Using $\mu(B_i) = \mu(C_i)$ for all $i \neq i_0$ (by (iii) and the continuity of μ) we then get

$$\begin{aligned} \mu^0(E) &= \mu^n(A) \\ &\geq \prod_{i \neq i_0} \mu(C_i)(\mu(B_{i_0}) + \mu(C_{i_0})) \\ &= \mu(B) + \mu(C) \\ &= \mu^0(E_1) + \mu^0(E_2) \end{aligned}$$

Similarly, by the right inequality of (2.13), it is shown that $\mu^0(E) \leq \mu^0(E_1) + \mu^0(E_2)$, and hence the additivity of μ^0.

We now turn to the proof of (i)-(iv). We have that

$$(\times B_i \cap X(J)) \;\cap\; (\times C_i \cap X(J)) \;=\; \times(B_i \cap C_i) \cap X(J) = \emptyset.$$

Thus, it must be true that for at least one $i_0 \in \{1, \dots, n\}$ the intersection $B_{i_0} \cap C_{i_0}$ is finite. A_{i_0} being infinite, one of B_{i_0} and C_{i_0} must be infinite too, meaning that at least one of $B_{i_0} \setminus C_{i_0}$ and $C_{i_0} \setminus B_{i_0}$ is infinite. Without loss of generality, assume that this is $B_{i_0} \setminus C_{i_0}$.

To prove (i), let $i \neq i_0$ and $c \in C_i$ be given. Let $\boldsymbol{c} \in \times C_i \cap X(J)$ with $\boldsymbol{c}(i) = c$. There exists $\boldsymbol{b} \in \times B_i \cap X(J)$ with $\boldsymbol{b}(i_0) \in B_{i_0} \setminus (C_{i_0} \cup \boldsymbol{c})$. Since both $\boldsymbol{b}$ and $\boldsymbol{c}$ are in $A \cap X(J)$, the tuple $\boldsymbol{c}'$ defined by $\boldsymbol{c}'(i) = \boldsymbol{c}(i)$ for $i \neq i_0$ and $\boldsymbol{c}'(i_0) = \boldsymbol{b}(i_0)$ must also belong to $A \cap X(J)$. $\boldsymbol{c}'$ can not be in E_2, thus $\boldsymbol{c}' \in E_1$, and $c \in B_i$. This proves $C_i \subseteq B_i$ for all $i \neq i_0$.

But now $B_{i_0} \cap C_{i_0} = \emptyset$ must hold. Otherwise, for $\boldsymbol{c} \in E_2$ with $\boldsymbol{c}(i_0) \in B_{i_0} \cap C_{i_0}$ we would have $\boldsymbol{c} \in E_1$, a contradiction.

To see (iii), let $\boldsymbol{c} \in E_2$. If $B_i \setminus C_i$ is infinite for some $i \neq i_0$, then there exists $b \in B_i \setminus C_i$ with $b \neq \boldsymbol{c}(j)$ for $j = 1, \dots, n$. Using the same argument as above, we obtain that the tuple $\boldsymbol{c}'$ with $\boldsymbol{c}'(j) = \boldsymbol{c}(j)$ for $j \neq i$ and $\boldsymbol{c}'(i) = b$ belongs to E_1, a contradiction because $\boldsymbol{c}(i_0) \notin B_{i_0}$.

(iv) directly follows from (i),(iii), and the fact that not both B_i and C_i can be finite.

It remains to show (2.13). We only show the left inequality, the right one being proven by analogous arguments. According to our premise

$$A \cap X(J) = (B \cup C) \cap X(J).$$

By

$$\times B_i \cup \times C_i \supseteq \times_{i \neq i_0}(B_i \cap C_i) \times (B_{i_0} \cup C_{i_0}) = \times_{i \neq i_0} C_i \times (B_{i_0} \cup C_{i_0})$$

it follows that

$$A \cap X(J) \supseteq \times_{i \neq i_0} C_i \times (B_{i_0} \cup C_{i_0}) \cap X(J). \tag{2.14}$$

To show that this inequality remains valid without the relativization to $X(J)$, let $i \in \{1, \dots, n\}$ and $c \in C_i$ if $i \neq i_0$, otherwise $c \in B_{i_0} \cup C_{i_0}$. By the infinity of the C_i and $B_{i_0} \cup C_{i_0}$, there exists $\boldsymbol{c} \in \times_{i \neq i_0} C_i \times (B_{i_0} \cup C_{i_0}) \cap X(J)$ with $\boldsymbol{c}(i) = c$. (2.14) then yields $\boldsymbol{c} \in A$, so that $C_i \subseteq A_i$ $(i \neq i_0)$, $B_{i_0} \cup C_{i_0} \subseteq A_{i_0}$, and we obtain the left inequality of (2.13). □

Theorem 1.2.10 now justifies the following definition.

Definition 2.4.5 Let $\mathfrak{A}_n^{\mathrm{m}}$, $\mathfrak{E}_n$, μ_n^0 be as defined by (2.7), (2.11), and (2.12) respectively. The unique probability measure extending μ_n^0 to $\mathfrak{A}_n^{\mathrm{m}}$ is denoted μ_n^{m}.

2.4.2 Consistency Properties of μ_n^{m}

Lemma 2.4.6 Homogeneity holds for μ_n^{m} $(n \geq 1)$.

Proof: Let $A \cap X(J) \in \mathfrak{E}_n$, π be a permutation of $\{1, \dots, n\}$. Then

$$\mu_n^{\mathrm{m}}(\pi(A \cap X(J))) = \mu_n^{\mathrm{m}}(\pi(A) \cap \pi(X(J))) = \begin{cases} \mu^n(\pi(A)) & \text{if } \pi(X(J)) = X(J^{\neq}) \\ 0 & \text{else.} \end{cases}$$

Since $\mu^n(\pi(A)) = \mu^n(A)$, and $\pi(X(J)) = X(J^{\neq})$ iff $J = J^{\neq}$, this is equal to $\mu_n^0(A \cap X(J))$. Thus, μ_n^0 is invariant under permutations on $\mathfrak{E}_n$.

With $\pi(E_1 \cup E_2) = \pi(E_1) \cup \pi(E_2)$, this implies homogeneity of μ_n^{m} on $\mathfrak{A}_n^{\mathrm{m}}$. □

Lemma 2.4.7 $(\mathfrak{A}_n^{\mathrm{m}})_n$ is closed under products.

Proof: Because of the distributivity of unions over products, i.e.

$$\bigcup_{i\in I} A_i \times \bigcup_{j\in J} B_j \;=\; \bigcup_{i\in I, j\in J} A_i \times B_i,$$

it suffices to show that the product of $E_1 \in \mathfrak{E}_k$ and $E_2 \in \mathfrak{E}_l$ is in $\mathfrak{A}_{k+l}^{\mathrm{m}}$.

With $E_1 = \times_{i=1}^k A_i \cap X(J),\ E_2 = \times_{j=1}^l B_j \cap X(J')$ we get

$$E_1 \times E_2 = \bigcup_{J^*} \Big[\Big(\times_{i=1}^k A_i \times \times_{j=1}^l B_j \Big) \cap X(J^*) \Big] \;\in\; \mathfrak{A}_{k+l}^{\mathrm{m}} \tag{2.15}$$

with J^* ranging over all equivalence relations on $\{1, \ldots, k+l\}$ that satisfy $(i,j) \in J^* \Leftrightarrow (i,j) \in J$ for $i,j \in \{1,\ldots,k\}$, and $(i,j) \in J^* \Leftrightarrow (i-k, j-k) \in J'$ for $i,j \in \{k+1, \ldots, k+l\}$ □

We now turn to the Fubini-property. For a motivation of how the proof that $(\mathfrak{A}_n^{\mathrm{m}}, \mu_n^{\mathrm{m}})_n$ has this property will proceed, consider the simpler case of the sequence of product algebras and product measures $(\mathfrak{A}^n, \mu^n)_n$.

Let $A = \times A_i \in \mathfrak{E}^{\times}$ be one of the measurable rectangles that generate $\mathfrak{A}^n$. To show that the Fubini-property holds for A, let $I \subset \{1, \ldots, n\}$ with $1 \leq k :=\mid I \mid < n$ and $I' := \{1, \ldots, n\} \setminus I$. For every $\boldsymbol{a} \in M^I$ we have $\sigma_{\boldsymbol{a}}^I(A) = \times_{j\in I'} A_j$ if $\boldsymbol{a}$ is in the projection $\rho^I(A) = \times_{i\in I} A_i$ of A onto M^I, and $\sigma_{\boldsymbol{a}}^I(A) = \emptyset$ else (cf. figure 2.3). Thus, for $r \in [0,1]$, $A_{I,r} = \rho^I(A)$ if $r \leq \mu_{n-k}(\times_{j\in I'} A_j)$, and $A_{I,r} = \emptyset$ else. With $\mu^n(A) = \mu^k(\times_{i\in I} A_i)\mu^{n-k}(\times_{j\in I'} A_j)$, this immediately establishes the Fubini-property for A.

This result for $A \in \mathfrak{E}^{\times}$ must then be generalized to arbitrary $A \in \mathfrak{A}^n$.

For $(\mathfrak{A}_n^{\mathrm{m}}, \mu_n^{\mathrm{m}})_n$ we will repeat essentially the same argument: first considering only elements $E \in \mathfrak{E}_n$ it is shown that the measure of sections $\sigma_{\boldsymbol{a}}^I(E)$ of E is independent of $\boldsymbol{a}$, provided $\boldsymbol{a}$ is in the projection of E onto M^I. To this end, lemma 2.4.8 provides an explicit representation of projections and sections of E, which enables us to prove in lemma 2.4.9 that $E \in \mathfrak{E}_n$ behaves just like a rectangle $A \in \mathfrak{E}^{\times}$ with respect to measures of sections and projections. It will then be easy to prove the Fubini-property for $(\mathfrak{A}_n^{\mathrm{m}}, \mu_n^{\mathrm{m}})_n$ in lemma 2.4.10.

Lemma 2.4.8 Let $E = \times A_i \cap X(J) \in \mathfrak{E}_n$, $E \neq \emptyset$, $I \subset \{1, \ldots, n\}$ with $1 \leq k :=\mid I \mid < n$, and $I' := \{1, \ldots, n\} \setminus I$. Then

(i) $\rho^I(E) = \times_{i\in I} A_i \cap X(J \restriction I)$.

(ii) For $\boldsymbol{a} \in \rho^I(E):\ \sigma_{\boldsymbol{a}}^I(E) = \times_{j\in I'} A_j^*(\boldsymbol{a}) \cap X(J \restriction I')$ with

$$A_j^*(\boldsymbol{a}) = \begin{cases} \{\boldsymbol{a}(i)\} & \text{if } \exists i \in I\ (i,j) \in J \\ A_j \setminus \{\boldsymbol{a}(i) \mid i \in I\} & \text{else.} \end{cases}$$

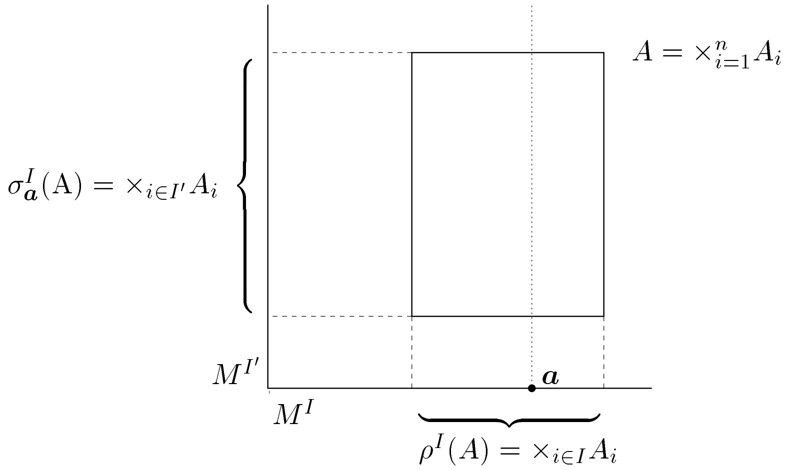

Figure 2.3: The Fubini property for measurable rectangles

Proof: Denote $\times_{i\in I} A_i \cap X(J\restriction I)$ by $E^{(i)}$, and $\times_{j\in I'} A^*_j(\boldsymbol{a}) \cap X(J\restriction I')$ by $E^{(ii)}(\boldsymbol{a})$. To prove the two parts of the lemma, we show that for all $\boldsymbol{a} \in M^I$:

$$\boldsymbol{a} \in E^{(i)} \;\Rightarrow\; E^{(ii)}(\boldsymbol{a}) \neq \emptyset, \tag{2.16}$$

and that for $\boldsymbol{a} \in M^I$, $\boldsymbol{b} \in M^{I'}$:

$$(\boldsymbol{a}, \boldsymbol{b}) \in E \;\text{ iff }\; \boldsymbol{a} \in E^{(i)} \text{ and } \boldsymbol{b} \in E^{(ii)}(\boldsymbol{a}). \tag{2.17}$$

With (2.16) and (2.17) part (i) of the lemma follows by the equivalences for $\boldsymbol{a} \in M^I$:

$$\begin{aligned}
\boldsymbol{a} \in \rho^I(E) \quad &\Leftrightarrow \exists \boldsymbol{b} \in M^{I'}\ (\boldsymbol{a}, \boldsymbol{b}) \in E \\
&\Leftrightarrow\ \boldsymbol{a} \in E^{(i)} \text{ and } E^{(ii)}(\boldsymbol{a}) \neq \emptyset \quad \text{(by (2.17))} \\
&\Leftrightarrow\ \boldsymbol{a} \in E^{(i)} \qquad\qquad\qquad\qquad \text{(by (2.16))}.
\end{aligned}$$

By definition, $\sigma^I_{\boldsymbol{a}}(E) = \{\boldsymbol{b} \mid (\boldsymbol{a}, \boldsymbol{b}) \in E\}$. For $\boldsymbol{a} \in \rho^I(E)$, i.e., by part (i), $\boldsymbol{a} \in E^{(i)}$, this set is equal to $E^{(ii)}(\boldsymbol{a})$ by (2.17), proving the second part of the lemma. It remains to show (2.16) and (2.17).

For the proof of (2.16), let $\boldsymbol{a} \in E^{(i)}$. To define an element $\boldsymbol{b} \in E^{(ii)}(\boldsymbol{a})$ let $I' = \{j_1, \ldots, j_{n-k}\}$, and define

$$\boldsymbol{b}(j_l) = \begin{cases} a & \text{if } A_{j_l} = \{a\} \\ \boldsymbol{a}(i) & \text{if } \exists i \in I\ (i, j_l) \in J \\ \boldsymbol{b}(j_h) & \text{if } \exists h < l\ (j_h, j_l) \in J \\ b(l) & \text{else} \end{cases}$$

with $b(l)$ an arbitrary element of the infinite set $A^*_{j_l}(\boldsymbol{a}) \setminus \{\boldsymbol{b}(j_1), \ldots, \boldsymbol{b}(j_{l-1})\}$.

By an induction on l $(1 \le l \le n-k)$, we show that $\boldsymbol{b}(j_l)$ is well-defined, i.e. when more than one of the first three cases distinguished in its definition holds, each of them defines the same element, that $\boldsymbol{b}(j_l) \in A^*_{j_l}(\boldsymbol{a})$, and that $(\boldsymbol{a}, (\boldsymbol{b}(j_1), \ldots, \boldsymbol{b}(j_l))) \in X(J \upharpoonright (I \cup \{j_1, \ldots, j_l\}))$ (this last assertion is a little stronger than needed to show that $\boldsymbol{b} \in E^{(ii)}(\boldsymbol{a})$, but is more convenient for the induction step).

The base case $l = 1$ of the induction is just a simpler variation of the induction step and is here omitted. Therefore, let $1 < l \leq n - k$.

Suppose that $A_{j_l} = \{a\}$ and $(i, j_l) \in J$ for some $i \in I$. Then, by the definition of $\mathfrak{C}_n$, $A_i = \{a\}$, and, because $\boldsymbol{a} \in E^{(i)}$, $\boldsymbol{a}(i) = a$. Similarly in the case $A_{j_l} = \{a\}$ and $(j_h, j_l) \in J$ for some $h < l$: again we then have that $A_{j_h} = \{a\} = A^*_{j_h}$ (using (2.10) for this last identity). With the induction hypothesis $\boldsymbol{b}(j_h) \in A^*_{j_h}$ we get $\boldsymbol{b}(j_h) = a$. The final case we have to consider when looking for a potential inconsistency in the definition of $\boldsymbol{b}(j_l)$ is when both $(i, j_l) \in J$ and $(i_h, i_l) \in J$. But in that case we also have $(i, j_h) \in J$, and by the induction hypothesis $\boldsymbol{b}(j_h)$ has been (well-) defined as $\boldsymbol{a}(i)$. Again, the two definitions of $\boldsymbol{b}(j_l)$ coincide. This completes the first part of the induction step: the definition of $\boldsymbol{b}(j_l)$ is consistent.

The second assertion that we have made is that $\boldsymbol{b}(j_l) \in A^*_{j_l}(\boldsymbol{a})$. This is immediately seen to be true if $A_{j_l} = \{a\}(= A^*_{j_l}(\boldsymbol{a}))$, or there exists $i \in I$ with $(i, j_l) \in J$. The third case, $\boldsymbol{b}(j_l)=$ $\boldsymbol{b}(j_h)$, is covered by $A^*_{j_h}(\boldsymbol{a}) = A^*_{j_l}(\boldsymbol{a})$ and the induction hypothesis $\boldsymbol{b}(j_h) \in A^*_{j_h}(\boldsymbol{a})$. When $\boldsymbol{b}(j_l)$ is defined by the "else"-case, $\boldsymbol{b}(j_l) \in A^*_{j_l}(\boldsymbol{a})$ trivially holds.

Finally, it must be ascertained that for all $h < l$: $\boldsymbol{b}(j_h) = \boldsymbol{b}(j_l) \Leftrightarrow (j_h, j_l) \in J$, and that for all $i \in I$: $\boldsymbol{a}(i) = \boldsymbol{b}(j_l) \Leftrightarrow (i, j_l) \in J$. The direction from right to left is trivial because of the definition of $\boldsymbol{b}(j_l)$. For the converse direction of the first equivalence assume that $(j_h, j_l) \notin J$. In the case $A_{j_l} = \{a\}$ this implies $A_{j_h} \cap A_{j_l} = \emptyset$ and thus $\boldsymbol{b}(j_h) \neq \boldsymbol{b}(j_l)$. In case that $\boldsymbol{b}(j_l)$ is defined as $\boldsymbol{a}(i)$ for $(i, j_l) \in J$, we get $(i, j_h) \notin J$, and by induction hypothesis $\boldsymbol{b}(j_h) \neq \boldsymbol{a}(i)$. The last case, $\boldsymbol{b}(j_l) = b(l)$ is trivial. For the left to right direction of the second equivalence assume that $(i, j_l) \notin J$. By analogous arguments as before, it is to shown that by whatever case $\boldsymbol{b}(j_l)$ has been defined, $\boldsymbol{a}(i) \neq \boldsymbol{b}(j_l)$ holds. This concludes the proof of (2.16).

The proof of (2.17) we begin with the left to right direction. Let $(\boldsymbol{a},\boldsymbol{b}) \in E$. Clearly, $\boldsymbol{a} \in X(J \upharpoonright I)$ and $\boldsymbol{b} \in X(J \upharpoonright I')$. It is immediate, too, that for all $i \in I,\ j \in I'$: $\boldsymbol{a}(i) \in A_i$, $\boldsymbol{b}(j) \in A^*_j(\boldsymbol{a})$, from which $\boldsymbol{a} \in E^{(i)}$, $\boldsymbol{b} \in E^{(ii)}(\boldsymbol{a})$ follows.

Conversely, let $\boldsymbol{a} \in E^{(i)}$, $\boldsymbol{b} \in E^{(ii)}(\boldsymbol{a})$. Obviously, then $(\boldsymbol{a},\boldsymbol{b}) \in \times_{i=1}^n A_i$. To show that $(\boldsymbol{a},\boldsymbol{b}) \in X(J)$ it is sufficient to remark that for $i \in I,\ j \in I'$ we have $\boldsymbol{a}(i) = \boldsymbol{b}(j)$ iff $(i, j) \in J$ by the definition of $A^*_j(\boldsymbol{a})$. □

With lemma 2.4.8 we can prove the next lemma, which essentially states that the Fubini-property holds for $E \in \mathfrak{C}_n$ in very much the same way as it is given for measurable rectangles (figure 2.3).

Lemma 2.4.9 Let E, I, I' as in lemma 2.4.8. Then $\rho^I(E) \in \mathfrak{A}^{\mathrm{m}}_k$, and for every $\boldsymbol{a} \in M^I$: $\sigma^I_{\boldsymbol{a}}(E) \in \mathfrak{A}^{\mathrm{m}}_{n-k}$. For $\boldsymbol{a},\boldsymbol{a}' \in \rho^I(E)$: $\mu^{\mathrm{m}}_{n-k}(\sigma^I_{\boldsymbol{a}}(E)) = \mu^{\mathrm{m}}_{n-k}(\sigma^I_{\boldsymbol{a}'}(E))$, and

$$\mu^{\mathrm{m}}_n(E) = \mu^{\mathrm{m}}_k(\rho^I(E))\mu^{\mathrm{m}}_{n-k}(\sigma^I_{\boldsymbol{a}}(E)). \tag{2.18}$$

Proof: The measurability of $\rho^I(E)$ and $\sigma^I_{\boldsymbol{a}}(E)$ is established by the representations given in lemma 2.4.8. Also, for $\boldsymbol{a} \in \rho^I(E)$ this lemma yields

$$\mu^{\mathrm{m}}_{n-k}(\sigma^I_{\boldsymbol{a}}(E)) = \begin{cases} 0 & \text{if } \exists i \in I, j \in I' \ (i,j) \in J, \\ & \text{ or } J \restriction I' \neq J^{\neq}_{n-k} \\ \prod_{j \in I'} \mu(A_j) & \text{else.} \end{cases} \tag{2.19}$$

Here it is used that when $A^*_j(\boldsymbol{a})$ is defined as $A_j \setminus \{\boldsymbol{a}(i) \mid i \in I\}$, $\mu(A^*_j(\boldsymbol{a})) = \mu(A_j)$ because μ is continuous. Particularly, (2.19) shows that $\mu_{n-k}(\sigma^I_{\boldsymbol{a}}(E))$ is independent of the particular choice of $\boldsymbol{a} \in \rho^I(E)$. To prove (2.18), first assume that $J \neq J^{\neq}_n$. Then $J \restriction I \neq J^{\neq}_k$, $J \restriction I' \neq J^{\neq}_{n-k}$, or there exists $i \in I$, $j \in I'$ with $(i,j) \in J$. In the first case we get that $\mu^{\mathrm{m}}_k(\rho^I(E)) = 0$, in the latter two cases that $\mu_{n-k}(\sigma^I_{\boldsymbol{a}}(E)) = 0$.

Now let $J = J^{\neq}_n$. Then $\mu^{\mathrm{m}}_k(\rho^I(E)) = \prod_{i \in I}(A_i)$. Together with (2.19) this yields

$$\mu^{\mathrm{m}}_n(E) = \prod_{i=1}^{n} \mu(A_i) = \mu^{\mathrm{m}}_k(\rho^I(E))\mu^{\mathrm{m}}_{n-k}(\sigma^I_{\boldsymbol{a}}(E)).$$

□

Lemma 2.4.10 $(\mathfrak{A}^{\mathrm{m}}_n, \mu^{\mathrm{m}}_n)_n$ has the Fubini-property.

Proof: Let $A \in \mathfrak{A}^{\mathrm{m}}_n$, I as in lemma 2.4.8. A has a representation $E_1 \dot\cup \ldots \dot\cup E_m$ with $E_i \in \mathfrak{E}_n$ $(i = 1, \ldots, m)$. For $\boldsymbol{a} \in M^I$ define

$$K(\boldsymbol{a}) := \{i \in \{1, \ldots, m\} \mid \boldsymbol{a} \in \rho^I(E_i)\}$$

(cf. figure 2.4(a)). Let $K_1, \ldots, K_l$ denote the elements of the set $\{K(\boldsymbol{a}) \mid \boldsymbol{a} \in M^I\}$. For $j \in \{1, \ldots, l\}$ let

$$M_j := \{\boldsymbol{a} \in M^I \mid K(\boldsymbol{a}) = K_j\} \ = \ \bigcap_{i \in K_j} \rho^I(E_i).$$

By lemma 2.4.8, the sets M_j are measurable. Since $\sigma^I_{\boldsymbol{a}}(\dot\cup E_i) \ = \ \dot\cup_{\sigma^I_{\boldsymbol{a}}}(E_i)$, we have for all $j \in \{1, \ldots, l\}$ and $\boldsymbol{a} \in M_j$

$$\mu^{\mathrm{m}}_{n-k}(\sigma^I_{\boldsymbol{a}}(A)) = \sum_{i \in K_j} \mu^{\mathrm{m}}_{n-k}(\sigma^I_{\boldsymbol{a}}(E_i)),$$

where, by lemma 2.4.9, this sum is independent of the particular choice of $\boldsymbol{a} \in M_j$.

Let $\boldsymbol{a}_j$ denote an arbitrary element of M_j. For $r \in [0, 1]$ define

$$J(r) := \left\{ j \in \{1, \ldots, l\} \mid \sum_{i \in K_j} \mu^{\mathrm{m}}_{n-k}(\sigma^I_{\boldsymbol{a}_j})(E_i)) \geq r \right\}.$$

(cf. figure 2.4(b)). Then

$$A_{I,r} = \dot{\bigcup_{j \in J(r)}} M_j,$$

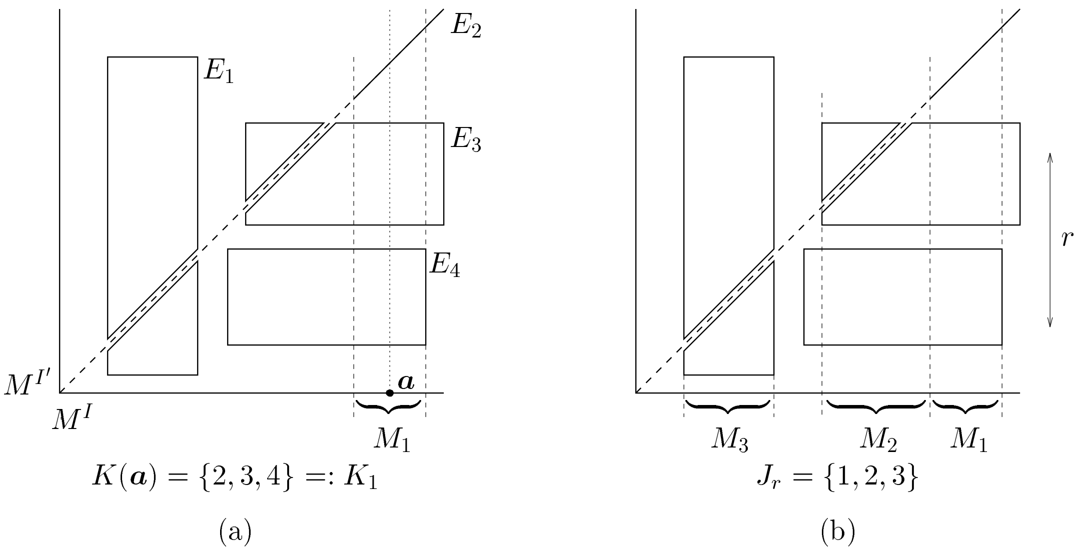

Figure 2.4: Proof of the Fubini-property

and

$$\begin{aligned} r\mu_k^{\mathrm{m}}(A_{I,r}) &= r\sum_{j\in J(r)} \mu_k^{\mathrm{m}}(M_j) \\ &\leq \sum_{j\in J(r)} [\sum_{i\in K_j} \mu_{n-k}^{\mathrm{m}}(\sigma_{\boldsymbol{a}_j}^I(E_i))]\mu_k^{\mathrm{m}}(M_j). \end{aligned} \tag{2.20}$$

Using

$$\mu_{n-k}^{\mathrm{m}}(\sigma_{\boldsymbol{a}_j}^I(E_i)) = \mu_{n-k}^{\mathrm{m}}(\sigma_{\boldsymbol{a}'_i}^I(E_i)),$$

with $\boldsymbol{a}'_i$ an arbitrary element in $\rho^I(E_i)$, the double sum in (2.20) can be rewritten as

$$\begin{aligned} &\sum_{i=1}^m \left[\mu_{n-k}^{\mathrm{m}}(\sigma_{\boldsymbol{a}'_i}^I(E_i)) \sum_{\{j\in J(r)\mid i\in K_j\}} \mu_k^{\mathrm{m}}(M_j)\right] \\ \leq\ &\sum_{i=1}^m [\mu_{n-k}^{\mathrm{m}}(\sigma_{\boldsymbol{a}'_i}^I(E_i))\mu_k^{\mathrm{m}}(\rho^I(E_i))] \\ =\ &\sum_{i=1}^m \mu_n^{\mathrm{m}}(E_i) \quad \text{(by lemma 2.4.9)} \\ =\ &\mu_n^{\mathrm{m}}(A). \end{aligned}$$

□

2.4.3 The Closure Property for $(\mathfrak{A}_n^{\mathrm{m}}, \mu_n^{\mathrm{m}})$

After the purely measure-theoretic considerations of the previous section, it now has to be shown that the measure algebras $(\mathfrak{A}_n^{\mathrm{m}}, \mu_n^{\mathrm{m}})$ are adequate for the language $\mathrm{L}_{\mathrm{S}}^{g}$ when S is monadic, i.e. satisfy the closure condition for such vocabularies.

The following lemma shows that this is the case, provided that the interpretations $I(\mathsf{R})$ of relation symbols $\mathsf{R} \in \mathrm{S}$ are in $\mathfrak{A}$. The lemma will be proven by an induction on the structure of ϕ, and therefore also contains a suitable auxiliary statement for field terms t.

Lemma 2.4.11 Let S be a monadic vocabulary, $(\mathfrak{A}, \mu)$ a continuous measure algebra over M, and I an interpretation function with $I(\mathsf{R}) \in \mathfrak{A}$ for $\mathsf{R} \in \mathrm{S}$. Then

(+) For every $\mathrm{t}(\boldsymbol{v},\boldsymbol{x}) \in \mathrm{FT}_{\mathrm{S}}^{g}$ there exists a finite partition $\{A_i \mid i = 1, \ldots, k\} \subseteq \mathfrak{A}_{|\boldsymbol{v}|}^{\mathrm{m}}$ of $M^{|\boldsymbol{v}|}$ such that for all $\boldsymbol{r} \in F^{|\boldsymbol{x}|}$, $i \in \{1, \ldots, k\}$, $\boldsymbol{a}, \boldsymbol{a}' \in A_i$:
$$(\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\mathrm{t}) = (\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}', \boldsymbol{x}/\boldsymbol{r})(\mathrm{t}).$$

(++) For all $\phi\langle \boldsymbol{v},\boldsymbol{x}\rangle \in \mathrm{L}_{\mathrm{S}}^{g}$ there exist $R_1, \ldots, R_m \subseteq F^{|\boldsymbol{x}|}$ and mutually disjoint $A_1, \ldots, A_m \in \mathfrak{A}_{|\boldsymbol{v}|}^{\mathrm{m}}$ with
$$\mathfrak{M}(\phi\langle \boldsymbol{v},\boldsymbol{x}\rangle) = \cup_{i=1}^{m}(A_i \times R_i).$$

Proof: We first note that it is sufficient to prove (++) without the condition of the A_i being disjoint, because for any $A_1, \ldots, A_m$, a representation with disjoint A_i' can be obtained via

$$\bigcup_{i=1}^{m}(A_i \times R_i) = \bigcup_{J \subseteq \{1,\ldots,m\}} \left[\Big(\bigcap_{i \in J} A_i \cap \bigcap_{i \notin J} A_i^c\Big) \times \bigcup_{i \in J} R_i \right].$$

The proof now follows the induction schema as described in 2.2.

(a): Let $\mathrm{t} \in \mathrm{FT}_{\mathrm{S}}^{g}$ be first-order. Then $\mathrm{t} \equiv \mathrm{t}(\boldsymbol{x})$ only contains field variables, and (+) trivially holds.

For first-order $\phi\langle \boldsymbol{v},\boldsymbol{x}\rangle \in \mathrm{L}_{\mathrm{S}}^{g}$ we prove (++) by induction on the structure of ϕ.
(aa): ϕ atomic: If ϕ is an atomic field formula, then ϕ does not contain any domain variables, and
$$\mathfrak{M}(\phi\langle \boldsymbol{x}\rangle) = R$$
for some $R \subseteq F^{|\boldsymbol{x}|}$. Atomic domain formulas ϕ can be either of the form $\mathsf{R}\mathrm{t}$, or $\mathrm{t}_1 = \mathrm{t}_2$ with domain terms $\mathrm{t}, \mathrm{t}_1, \mathrm{t}_2$. S being monadic, these terms are either constant or variable symbols.

Under the condition that ϕ has at least one free variable, this leaves us with three cases to be distinguished:

$$\begin{aligned}
\phi\langle v\rangle \equiv \mathsf{R}v \quad &\to \quad \mathfrak{M}(\phi\langle v\rangle) = I(\mathsf{R}) \in \mathfrak{A}_1^{\mathrm{m}} \\
\phi\langle v\rangle \equiv v = \mathsf{a} \quad &\to \quad \mathfrak{M}(\phi\langle v\rangle) = \{I\mathsf{a}\} \in \mathfrak{A}_1^{\mathrm{m}} \\
\phi\langle v, w\rangle \equiv v = w \quad &\to \quad \mathfrak{M}(\phi\langle v, w\rangle) = \{(a, a) \mid a \in M\} \in \mathfrak{A}_2^{\mathrm{m}}
\end{aligned}$$

(ab): Boolean operations: Let $\phi\langle \boldsymbol{v},\boldsymbol{x}\rangle \equiv \psi\langle \boldsymbol{v}',\boldsymbol{x}'\rangle \wedge \chi\langle \boldsymbol{v}'',\boldsymbol{x}''\rangle$. We may assume that $\boldsymbol{v} = \boldsymbol{v}' = \boldsymbol{v}''$ and $\boldsymbol{x} = \boldsymbol{x}' = \boldsymbol{x}''$ because if not, we can replace $\psi\langle \boldsymbol{v}',\boldsymbol{x}'\rangle$ by

$$\tilde{\psi}\langle \boldsymbol{v},\boldsymbol{x}\rangle :\equiv \psi\langle \boldsymbol{v}',\boldsymbol{x}'\rangle \wedge \bigwedge_{v\in \boldsymbol{v}\setminus\boldsymbol{v}'} v = v \wedge \bigwedge_{x\in\boldsymbol{x}\setminus\boldsymbol{x}'} x = x;$$

similarly for χ. By induction hypothesis

$$\mathfrak{M}(\psi\langle \boldsymbol{v},\boldsymbol{x}\rangle) = \cup_{i=1}^m A_i \times R_i, \quad \mathfrak{M}(\chi\langle \boldsymbol{v},\boldsymbol{x}\rangle) = \cup_{j=1}^l B_j \times S_j.$$

Then

$$\begin{aligned}\mathfrak{M}(\phi\langle \boldsymbol{v},\boldsymbol{x}\rangle) &= (\cup_{i=1}^m A_i \times R_i) \cap (\cup_{j=1}^l B_j \times S_j)\\ &= \cup_{i,j}(A_i \times R_i) \cap (B_j \times S_j)\\ &= \cup_{i,j}(A_i \cap B_j) \times (R_i \cap S_j).\end{aligned}$$

For $\phi\langle \boldsymbol{v},\boldsymbol{x}\rangle \equiv \neg\psi\langle \boldsymbol{v},\boldsymbol{x}\rangle$ with $\mathfrak{M}(\psi\langle \boldsymbol{v},\boldsymbol{x}\rangle) = \cup_{i=1}^m A_i \times R_i$, we get

$$\begin{aligned}\mathfrak{M}(\phi\langle \boldsymbol{v},\boldsymbol{x}\rangle) &= (\cup_i A_i \times R_i)^c\\ &= \cap_i (A_i \times R_i)^c\\ &= \cap_i[(A_i^c \times F^{|\boldsymbol{x}|}) \cup (M^{|\boldsymbol{v}|} \times R_i^c)].\end{aligned}$$

This set is as in (++) because each factor in the intersection has the prescribed form, and, as shown in the previous step, this form is preserved under finite intersections.
(ac): Quantification: Let $\phi\langle \boldsymbol{v},\boldsymbol{x}\rangle \equiv \exists u \psi\langle \boldsymbol{v}',\boldsymbol{x}\rangle$, where $|\boldsymbol{v}'| = |\boldsymbol{v}| + 1$, $\boldsymbol{v}'(j) = u$ for some $j \leq |\boldsymbol{v}'|$, $\boldsymbol{v}'(i) = \boldsymbol{v}(i)$ for $i < j$, $\boldsymbol{v}'(i) = \boldsymbol{v}(i-1)$ for $i > j$. By induction hypothesis $\mathfrak{M}(\psi\langle \boldsymbol{v}',\boldsymbol{x}\rangle) = \cup_i(A_i \times R_i)$ with $A_i \in \mathfrak{A}^{\mathrm{m}}_{|\boldsymbol{v}|+1}$. $\mathfrak{M}(\phi\langle \boldsymbol{v},\boldsymbol{x}\rangle)$ is the projection of $\mathfrak{M}(\psi\langle \boldsymbol{v}',\boldsymbol{x}\rangle)$ onto the coordinates distinct from the j-th domain coordinate. Denote this projection by ρ. Then

$$\begin{aligned}\rho(\cup_i(A_i \times R_i)) &= \cup_i\ \rho(A_i \times R_i)\\ &= \cup_i\ \rho^{\{1,\ldots,|\boldsymbol{v}'|\}\setminus\{j\}}(A_i) \times R_i.\end{aligned}$$

By lemma 2.4.8, the given projection of A_i is in $\mathfrak{A}^{\mathrm{m}}_{|\boldsymbol{v}|}$, so that the result again is of the specified form. The argument for the quantification of field variables is analogous.

(b): Let $\mathrm{t}(\boldsymbol{v},\boldsymbol{x}) \in \mathrm{FT}^g_{\mathrm{S}}$. For $\mathrm{t} \equiv x$, $\mathrm{t} \equiv 0$ and $\mathrm{t} \equiv 1$ (+) has been shown in (a). Let $\mathrm{t} \equiv \mathrm{t}_1 \cdot \mathrm{t}_2$, $\{A_i \mid i = 1,\ldots,k_1\}$ the partition of $M^{|\boldsymbol{v}|}$ associated with t_1, and $\{B_j \mid j = 1,\ldots,k_2\}$ the partition associated with t_2. Then, for all $\boldsymbol{r} \in F^{|\boldsymbol{x}|}$, $i \in \{1,\ldots,k_1\}, j \in \{1,\ldots,k_2\}$, $(\mathfrak{M},\boldsymbol{v}/\boldsymbol{a},\boldsymbol{x}/\boldsymbol{r})(\mathrm{t}_1 \cdot \mathrm{t}_2)$ is constant for all $\boldsymbol{a} \in A_i \cap B_j$. The system $\{A_i \cap B_j \mid i = 1,\ldots,k_1, j = 1,\ldots,k_2\}$ therefore is a partition of $\mathfrak{A}^{\mathrm{m}}_{|\boldsymbol{v}|}$ for which (+) holds for t. Analogously for $\mathrm{t}_1 + \mathrm{t}_2$.

Let $\mathrm{t}(\boldsymbol{v},\boldsymbol{x}) \equiv [\phi\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle]_{\boldsymbol{w}}$. By (++), $\mathfrak{M}(\phi\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle)$ has a representation $\cup_{i=1}^m(A_i \times R_i)$ with mutually disjoint $A_i \in \mathfrak{A}^{\mathrm{m}}_{|\boldsymbol{v}|+|\boldsymbol{w}|}$, $R_i \subseteq F^{|\boldsymbol{x}|}$. We may assume that each A_i is in fact an element of $\mathfrak{E}_{|\boldsymbol{v}|+|\boldsymbol{w}|}$: otherwise replace A_i by its representation as the disjoint finite union of elements from the generating system. Let $\boldsymbol{r} \in F^{|\boldsymbol{x}|}$, $\boldsymbol{a} \in M^{|\boldsymbol{v}|}$, and $\sigma_{\boldsymbol{a},\boldsymbol{r}}(\mathfrak{M}(\phi\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle))$ the section of $\mathfrak{M}(\phi\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle)$ along $\boldsymbol{a}$ and $\boldsymbol{r}$ at the coordinates of $\boldsymbol{v}$ and $\boldsymbol{x}$. Then

$$\begin{aligned}(\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\mathrm{t}) &= \mu^{\mathrm{m}}_{|\boldsymbol{w}|}((\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\phi\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle)) \\ &= \mu^{\mathrm{m}}_{|\boldsymbol{w}|}(\sigma_{\boldsymbol{a},\boldsymbol{r}}(\mathfrak{M}(\phi(\boldsymbol{v},\boldsymbol{w},\boldsymbol{x}))) \\ &= \mu^{\mathrm{m}}_{|\boldsymbol{w}|}(\bigcup_{i=1}^{m} \sigma_{\boldsymbol{a},\boldsymbol{r}}(A_i \times R_i)) \qquad (2.21)\end{aligned}$$

For each i, we have $\sigma_{\boldsymbol{a},\boldsymbol{r}}(A_i \times R_i) = \emptyset$ it $\boldsymbol{r} \notin R_i$, and $\sigma_{\boldsymbol{a},\boldsymbol{r}}(A_i \times R_i) = \sigma_{\boldsymbol{a}}(A_i)$ (the section of A_i along $\boldsymbol{a}$ at the coordinates of $\boldsymbol{v}$) else. Hence, the expression (2.21) is equal to

$$\mu^{\mathrm{m}}_{|\boldsymbol{w}|}(\bigcup_{\substack{i=1\\ \boldsymbol{r}\in R_i}}^{m} \sigma_{\boldsymbol{a}}(A_i)) = \sum_{\substack{i=1\\ \boldsymbol{r}\in R_i}}^{m} \mu^{\mathrm{m}}_{|\boldsymbol{w}|}(\sigma_{\boldsymbol{a}}(A_i)).$$

The last equality is due to the fact that with the A_i being disjoint, so are the section $\sigma_{\boldsymbol{a}}(A_i)$.

By lemma 2.4.9 the measure of $\sigma_{\boldsymbol{a}}(A_i)$ only depends on whether $\boldsymbol{a}$ is in the projection $\rho(A_i)$ of A_i onto the coordinates of $\boldsymbol{v}$. Hence, $(\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\mathrm{t}) = (\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}', \boldsymbol{x}/\boldsymbol{r})(\mathrm{t})$, if

$$\forall i \in 1, \dots, m \quad \boldsymbol{a} \in \rho(A_i) \Leftrightarrow \boldsymbol{a}' \in \rho(A_i)$$

This condition defines a finite partition of $M^{|\boldsymbol{v}|}$, thus proving (+).

(c): Let $\phi\langle \boldsymbol{v},\boldsymbol{x}\rangle \equiv \mathrm{t}_1(\boldsymbol{v},\boldsymbol{x}) \leq \mathrm{t}_2(\boldsymbol{v},\boldsymbol{x})$, $\{A_i \mid i = 1,\dots,k_1\}$ the partition of $M^{|\boldsymbol{v}|}$ associated with t_1, and $\{B_j \mid j = 1,\dots,k_2\}$ the partition associated with t_2. For $i \in \{1,\dots,k_1\}$, $j \in \{1,\dots,k_2\}$, $\boldsymbol{a},\boldsymbol{a}' \in A_i \cap B_j$, $\boldsymbol{r} \in F^{|\boldsymbol{x}|}$ then

$$(\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r}) \models \phi\langle \boldsymbol{v},\boldsymbol{x}\rangle \;\Leftrightarrow\; (\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}', \boldsymbol{x}/\boldsymbol{r}) \models \phi\langle \boldsymbol{v},\boldsymbol{x}\rangle,$$

so that we can define

$$R_{i,j} := \{\boldsymbol{r} \in F^{|\boldsymbol{x}|} \mid (\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r}) \models \phi\langle \boldsymbol{v},\boldsymbol{x}\rangle \text{ for } \boldsymbol{a} \in A_i \cap B_j\},$$

and obtain

$$\mathfrak{M}(\phi\langle \boldsymbol{v},\boldsymbol{x}\rangle) = \bigcup_{i,j}(A_i \cap B_j) \times R_{i,j}.$$

All other induction steps for ϕ are as in (a). $\square$

Theorem 2.4.12 Let $(\mathfrak{A}, \mu)$ be a continuous $\mathfrak{F}$-measure algebra over M. Let S be a monadic vocabulary, and $I : \mathrm{S} \to \mathfrak{A}$. Then $(M, I, \mathfrak{F}, (\mathfrak{A}^{\mathrm{m}}_n, \mu^{\mathrm{m}}_n)_n)$ is a statistical S-structure.

Proof: Homogeneity, product-, and Fubini-property have been proven in section 2.4.2. It remains to show that $(\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\phi\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle) \in \mathfrak{A}^{\mathrm{m}}_{|\boldsymbol{w}|}$ for all $\phi\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle \in \mathrm{L}^{g}_{\mathrm{S}}$, $\boldsymbol{a} \in M^{\boldsymbol{v}}$, $\boldsymbol{r} \in F^{\boldsymbol{x}}$. This, however, is immediate from (++) of lemma 2.4.11. By that lemma we have

$$\mathfrak{M}(\phi\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle) = \bigcup_{i=1}^{m} A_i \times R_i$$

with $A_i \in \mathfrak{A}^{\mathrm{m}}_{|\boldsymbol{v}|+|\boldsymbol{w}|}$. $(\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\phi\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle)$ thus is the union of sections $\sigma_{\boldsymbol{a}}(A_i)$, which by lemma 2.4.8 are in $\mathfrak{A}^{\mathrm{m}}_{|\boldsymbol{w}|}$. □

With the class of real-discrete structures (example 2.3.7) and the class of continuous structures for monadic languages, as given by theorem 2.4.12, we now have provided constructive descriptions of two distinct types of statistical S-structures. The lengthiness of the discussion that has been necessary to establish that a sequence of measure algebras of as simple a form as $(\mathfrak{A}^{\mathrm{m}}_n, \mu^{\mathrm{m}}_n)_n$ satisfies the consistency and closure conditions is an indication for how difficult it might be, in general, to construct statistical S-structures. Future enquiries into the model theory of $\mathscr{L}^\sigma$, therefore, should be directed at obtaining further results on the existence of statistical S-structures.

Of particular interest would be a general theorem stating under what conditions a measure algebra $(\mathfrak{A}, \mu)$ over M can be embedded into the one-dimensional algebra $(\mathfrak{A}_1, \mu_1)$ of a statistical S-structure.

To illustrate the need for such a theorem, consider the following L^σ-sentence:

$$\phi^{\mathrm{inf}} :\equiv \exists u \Big([w = u]_w > 0 \wedge \exists v (v \neq u \wedge [w = v]_w = \frac{1}{2}) \wedge \\ \forall v (v \neq u \rightarrow \exists v' (v' \neq u \wedge [w = v']_w = \frac{1}{2}[w = v]_w)) \Big).$$

This sentence (in a similar fashion as Abadi and Halpern's [1989] encoding of arithmetic in (real-valued) probability logic) postulates the existence of a sequence of domain elements $a_1, a_2, \ldots$ with statistical probabilities $\mu(\{a_i\}) = 1/2^i$. Furthermore, there is one element $b \neq a_i$ $(i \geq 1)$ that also has a positive probability. Clearly, this sentence is not satisfiable by real-valued probabilities. Taking $\mathfrak{F} = \mathfrak{R}^*$ (cf. example 1.2.4), however, we can construct a measure algebra $(\mathfrak{A}, \mu)$ over $M = \{b, a_1, a_2, \ldots\}$ that is consistent with ϕ^{inf} by letting $\mathfrak{A} = \mathfrak{A}^{\mathrm{c/f}}$ the algebra of finite and co-finite subsets of M, and μ on $\mathfrak{A}^{\mathrm{c/f}}$ be defined by $\mu(\{a_i\}) = 1/2^i$, $\mu(\{b\}) = \epsilon$, with ϵ an infinitesimal.

To prove that ϕ^{inf} is satisfiable in $\mathscr{L}^\sigma$ (as seems save to conjecture it is), it now has to be shown that there exists a statistical S-structure in which $(\mathfrak{A}^{\mathrm{c/f}}, \mu) \subseteq (\mathfrak{A}_1, \mu_1)$. Such a structure can neither be real-discrete nor continuous, and hence must be obtained by other methods than have yet been supplied. A general theorem on the embeddability of an algebra $(\mathfrak{A}, \mu)$ in a statistical S-structure might be a powerful tool to prove, in cases like this, the consistency of L^σ-formulas.

2.4.4 An Example

By a slight modification of the construction of μ^{m}_n, we can now give an example of a sequence $(\mathfrak{A}_n, \mu_n)_n$ that satisfies Homogeneity and the product property, but fails to have the Fubini property. This failure will be of a nontrivial sort in that the measurability conditions (2.2) and (2.3) are satisfied, but (2.4) is violated.

Example 2.4.13 Let $M = \mathbf{N}$, $\mathfrak{A} = \mathfrak{A}^{\mathrm{c/f}}$, $\mu = \mu^{\mathrm{c/f}}$ (cf. example 1.2.2 and 1.2.8). Let $\mathfrak{A}_n^{\mathrm{m}}$, $\mathfrak{E}_n$ be defined as in section 2.4.1. Let $J_n^{=}$ be the universal equivalence relation on $\{1, \ldots, n\}$, i.e. $J_n^{=} = \{(i,j) \mid i,j \in \{1, \ldots, n\}\}$, so that

$$X(J_n^{=}) = \{\boldsymbol{a} \in M^n \mid \boldsymbol{a}(i) = \boldsymbol{a}(j)\ \forall i,j\}.$$

Define on $\mathfrak{E}_n \setminus \{\emptyset\}$

$$\bar{\mu}_n^0(A \cap X(J)) = \begin{cases} \mu^n(A) & \text{if } J = J_n^{=} \\ 0 & \text{else,} \end{cases} \tag{2.22}$$

and let $\bar{\mu}_n^0(\emptyset) = 0$. We now show that we can obtain measures $\bar{\mu}_n$ from $\bar{\mu}_n^0$ that satisfy homogeneity and the product property.

As for μ_n^0, it must first be confirmed that $\bar{\mu}_n^0$ is additive: let $E = A \cap X(J)$, $E_1 = B \cap X(J')$, $E_2 = C \cap X(J'') \in \mathfrak{E}_n \setminus \emptyset$. $E = E_1 \dot{\cup} E_2$. Again we have $J = J' = J''$, and additivity trivially holds if $J \neq J_n^{=}$.

In the case $J = J_n^{=}$, by the definition of $\mathfrak{E}_n$, we have that $A_i = A_j$, $B_i = B_j$, $C_i = C_j$ for all i, j. A_1 is either a singleton, or a cofinite set. In the first case E itself is a singleton, and can in fact not be the union of two nonempty sets. Thus, A_1 is cofinite, and either B_1 or C_1 is cofinite also. Assume this is B_1. By the disjointness of E_1 and E_2 this implies that C_1 is a singleton, so that $\bar{\mu}_n^0(E) = \bar{\mu}_n^0(E_1) = 1$, $\bar{\mu}_n^0(E_2) = 0$. This proves the additivity of $\bar{\mu}_n^0$.

Let $\bar{\mu}_n$ be the extension of $\bar{\mu}_n^0$ to a probability measure on $\mathfrak{A}_n^{\mathrm{m}}$.

$X(J_n^{=})$ being invariant under permutations, homogeneity is shown for $\bar{\mu}_n$ as in the proof of lemma 2.4.6.

The product property only has to be shown for elements of the generating systems: let $E_1 = A \cap X(J) \in \mathfrak{E}_k$, $E_2 = B \cap X(J') \in \mathfrak{E}_l$. The product $E_1 \times E_2$, is again given by (2.15). If $J \neq J_k^{=}$, or $J' \neq J_l^{=}$, then $J^* \neq J_{k+l}^{=}$ for all J^*, so that $\bar{\mu}_{k+l}(E_1 \times E_2) = \bar{\mu}_k(E_1)\bar{\mu}_l(E_2) = 0$, which is also true when $J = J_k^{=}$, $J' = J_l^{=}$, but A_1 or B_1 is a singleton. When $J = J_k^{=}$, $J' = J_l^{=}$, and both A_1 and B_1 are cofinite, then so is $C := A_1 \cap B_1$, and from

$$E_1 \times E_2 \subseteq \times_{i=1}^{k+l} C \cap X(J_{k+l}^{=})$$

it follows that $\bar{\mu}_{k+l}(E_1 \times E_2) = \bar{\mu}_k(E_1)\bar{\mu}_l(E_2) = 1$.

That $(\mathfrak{A}_n^{\mathrm{m}}, \bar{\mu}_n)$ satisfies the measurability condition (2.2) is immediate from lemma 2.4.8 (ii), which is independent of the definition of a measure on $\mathfrak{A}_n^{\mathrm{m}}$. The measurability conditions (2.3) also can be proven as in lemma 2.4.10. However, from example 2.3.11 we know that $(\mathfrak{A}_n^{\mathrm{m}}, \bar{\mu}_n)$ does not have the Fubini property because $\bar{\mu}_n$ concentrates all probability mass on "hyperplanes".

2.5 $\mathscr{L}^\sigma$ Is First-Order Logic

By lemma 2.3.13, the logic $\mathscr{L}^\sigma$ is an extension of first-order logic. Since the syntax rules of L^σ strictly extend the syntax rules of first-order logic, this extension, in a sense, is strict. In this section it will be shown that the additional syntactic construct of statistical quantifiers can be represented in standard first-order syntax, and that $\mathscr{L}^\sigma$, in fact, can be understood as a special formalism within first-order logic. A completeness result for $\mathscr{L}^\sigma$then follows.

To reduce $\mathscr{L}^\sigma$ to first-order logic, first a translation from L_S^g to the first-order language $\mathrm{L}_{\mathrm{S}_\infty}$ in a vocabulary $\mathrm{S}_\infty \supset \mathrm{S}$ is defined. Then it is shown that the statistical S-structures correspond to the class of standard model theoretic S_∞- structures satisfying a certain set of axioms, such that

$$\mathfrak{M} \models_\sigma \phi \quad \text{iff} \quad \mathfrak{M}^* \models \phi^*,$$

where $\mathfrak{M}$ is a statistical S-structure, $\phi \in \mathrm{L}_\mathrm{S}^g$, $\mathfrak{M}^*$ is the S_∞- structure corresponding to $\mathfrak{M}$, and ϕ^* is the translation of ϕ.

The method that will be used to achieve this result is related to the technique of skolemization. Just as quantifiers are replaced by function symbols in skolemization, we will use function symbols to eliminate the statistical quantifier $[\;]_{\boldsymbol{w}}$.

2.5.1 Substitution

The translations that we will be defining require that we first touch upon one of the most dull of technicalities: substitution. Our treatment of this subject here is very much the same as the one given in [Bacchus, 1990a].

In order to handle substitution properly, a set of rules must be given that for a formula $\phi(\boldsymbol{v},\boldsymbol{x})$, domain terms $\mathbf{t}$, and field terms $\mathbf{s}$ defines the formula $\phi(\boldsymbol{v},\boldsymbol{x})[\boldsymbol{v}/\mathbf{t},\boldsymbol{x}/\mathbf{s}]$ in which the variables $\boldsymbol{v}$ and $\boldsymbol{x}$ are replaced by t and s respectively, and the bound variables of ϕ are appropriately renamed.

Substitution in $\phi \in \mathrm{L}_\mathrm{S}^g$ is defined by induction on the construction of ϕ just as in standard first-order logic, with the following additional rule for substituting into a probability term.

Definition 2.5.1 Let $\phi(\boldsymbol{v},\boldsymbol{w},\boldsymbol{x}) \in \mathrm{L}_\mathrm{S}^g$, $\mathbf{t}(\boldsymbol{v}')$ a tuple of domain-terms, and $\mathbf{s}(\boldsymbol{v}',\boldsymbol{x}')$ a tuple of field-terms with $|\,\mathbf{t}\,|=|\,\boldsymbol{v}\,|$ and $|\,\mathbf{s}\,|=|\,\boldsymbol{x}\,|$. Assume that substitution in ϕ has been defined. Let $\boldsymbol{w}_\mathrm{new}$ be the $|\,\boldsymbol{w}\,|$ - tuple of domain variables that in its i-th place has $\boldsymbol{w}(i)$ if $\boldsymbol{w}(i) \notin \boldsymbol{v}'$, and a new variable, not occuring in either ϕ or $\boldsymbol{v}'$, else. Then

$$[\phi(\boldsymbol{v},\boldsymbol{w},\boldsymbol{x})]_{\boldsymbol{w}}\,[\boldsymbol{v}/\mathbf{t},\boldsymbol{x}/\mathbf{s}] := [\phi(\boldsymbol{v},\boldsymbol{w},\boldsymbol{x})[\boldsymbol{w}/\boldsymbol{w}_\mathrm{new}][\boldsymbol{v}/\mathbf{t},\boldsymbol{x}/\mathbf{s}]]_{\boldsymbol{w}_\mathrm{new}}.$$

Example 2.5.2 To make the substitution $[v_3/\mathsf{g}(v_2)]$ in $\phi\langle v_3\rangle \equiv [\exists v_4\,\mathsf{f}(v_4,v_3,v_2)=v_1]_{(v_1,v_2)}$, define $\boldsymbol{w}_\mathrm{new} := (v_1,v_5)$. Then

$$\begin{aligned}\phi[v_3/\mathsf{g}(v_2)] \quad &\equiv [\phi[(v_1,v_2)/(v_1,v_5)][v_3/\mathsf{g}(v_2)]]_{(v_1,v_5)}\\ &\equiv [\exists v_4\,\mathsf{f}(v_4,\mathsf{g}(v_2),v_5)=v_1]_{(v_1,v_5)}.\end{aligned}$$

The following lemma concerns the associativity of two substitutions. It will be needed in the proof of some technical lemmas below.

Lemma 2.5.3 Let $\mathrm{t}(\boldsymbol{v},\boldsymbol{x}) \in \mathrm{FT}_\mathrm{S}^\sigma$, $\phi(\boldsymbol{v},\boldsymbol{x}) \in \mathrm{L}_\mathrm{S}^g$,

$$\begin{aligned}\mathbf{t}(\boldsymbol{v}') &:= (\mathrm{t}_1(\boldsymbol{v}'),\ldots,\mathrm{t}_{|\boldsymbol{v}|}(\boldsymbol{v}')) \subseteq \mathrm{DT}_\mathrm{S}^\sigma, \qquad \text{and}\\ \mathbf{s}(\boldsymbol{v}',\boldsymbol{x}') &:= (\mathrm{s}_1(\boldsymbol{v}',\boldsymbol{x}'),\ldots,\mathrm{s}_{|\boldsymbol{x}|}(\boldsymbol{v}',\boldsymbol{x}')) \subseteq \mathrm{FT}_\mathrm{S}^\sigma.\end{aligned}$$

Then

(+)
$$\models_\sigma \mathrm{t}[\boldsymbol{v}/\mathbf{t}(\boldsymbol{v}'), \boldsymbol{x}/\mathbf{s}(\boldsymbol{v}', \boldsymbol{x}')][\boldsymbol{v}'/\boldsymbol{v}'', \boldsymbol{x}'/\boldsymbol{x}''] = \mathrm{t}[\boldsymbol{v}/\mathbf{t}[\boldsymbol{v}'/\boldsymbol{v}''], \boldsymbol{x}/\mathbf{s}[\boldsymbol{v}'/\boldsymbol{v}'', \boldsymbol{x}'/\boldsymbol{x}'']].$$

(++)
$$\models_\sigma \phi[\boldsymbol{v}/\mathbf{t}(\boldsymbol{v}'), \boldsymbol{x}/\mathbf{s}(\boldsymbol{v}', \boldsymbol{x}')][\boldsymbol{v}'/\boldsymbol{v}'', \boldsymbol{x}'/\boldsymbol{x}''] \leftrightarrow \phi[\boldsymbol{v}/\mathbf{t}[\boldsymbol{v}'/\boldsymbol{v}''], \boldsymbol{x}/\mathbf{s}[\boldsymbol{v}'/\boldsymbol{v}'', \boldsymbol{x}'/\boldsymbol{x}'']].$$

Proof: The proof is by induction on the structure of t and ϕ, the details of which are omitted. Note, though, that for the base cases of the induction it is essential that for each of the variables of t (each of the variables of ϕ) a substituting term is explicitly specified – even though it may be the variable itself: consider the term $\mathrm{t} \equiv x$ as an example. Then

$$\mathrm{t}[y/x][x/1] \equiv \mathrm{t}[x/1] \equiv 1$$

while

$$\mathrm{t}[y/x[x/1]] \equiv \mathrm{t}[y/1] \equiv x,$$

clearly non-equivalent terms. On the other hand

$$\mathrm{t}[x/x, y/x][x/1] \equiv \mathrm{t}[x/1] \equiv 1$$

as before, but now

$$\mathrm{t}[x/x[x/1], y/x[x/1]] \equiv \mathrm{t}[x/1, y/1] \equiv 1$$

as well. □

2.5.2 The Translation

For a vocabulary S we define a vocabulary $\mathrm{S}_\infty \supset \mathrm{S}$, a translation $(\cdot)^* : \mathrm{L}^\sigma_\mathrm{S} \to \mathrm{L}_{\mathrm{S}_\infty}$, and a translation $(\cdot)^{-1} : \mathrm{L}_{\mathrm{S}_\infty} \to \mathrm{L}^\sigma_\mathrm{S}$.

For the sake of convenience, we shall work with a sorted first-order logic, i.e. just as in $\mathrm{L}^\sigma_\mathrm{S}$ we will use in the target language $\mathrm{L}_{\mathrm{S}_\infty}$ variables v and x designating objects of two different sorts (called D for domain and F for field). With each n-ary function symbol $\mathsf{f} \in \mathrm{S}_\infty$ a tuple $(\mathrm{s}_1, \ldots, \mathrm{s}_{n+1})$ ($\mathrm{s}_i \in \{\mathrm{D}, \mathrm{F}\}$) is associated, meaning that in a well formed formula the i-th argument of f must be of sort s_i ($i \leq n$), and the resulting f-term itself is of sort s_{n+1}. Similarly, an n-ary relation symbol R is said to be of sort $(\mathrm{s}_1, \ldots, \mathrm{s}_n)$ iff the i-th argument of R is of sort s_i.

Using variables of different sorts is not a true extension of first-order logic. The increase in expressive power obtained in this way could also be gained within the framework of standard first-order logic by introducing two new unary relation symbols D and F, and relativizing each quantification of a variable to the appropriate one.

The basic principle of our translation is best demonstrated by an example first. Consider the probability term

$$\mathrm{t}\langle u, v\rangle \equiv [\mathsf{R}uvw \wedge \exists v' \, \mathsf{S}v'w]_w.$$

In L^σ this term behaves like a binary function symbol with the two arguments u and v. We may therefore try to replace t by a standard first-order term

$$\mathsf{f}^l(u,v)$$

where f^l is a new function symbol with l a name or label encoding the fact that $\mathsf{f}^l(\cdot,\cdot)$ represents $[\mathsf{R}\cdot\cdot w \wedge \exists v'\,\mathsf{S}v'w]_w$.

This label will be of the form

$$l \equiv \mathsf{R}\alpha_1\alpha_2 w \wedge \exists v'\,\mathsf{S}v'w, w$$

with α_1, α_2 auxiliary variables standing for the place where the first and second argument of f^l are to be substituted. The term $\mathsf{f}^l(u,v)$ then contains all the information necessary to regain t.

The translation of t to $\mathsf{f}^l(u,v)$, however, is not yet uniquely determined: apart from the given function symbol f^l, we might also use the label

$$l' \equiv \mathsf{R}\alpha_2\alpha_1 w \wedge \exists v'\,\mathsf{S}v'w, w$$

where the two auxiliary variables α_1 and α_2 have been interchanged, and translate t by $\mathsf{f}^{l'}(v,u)$.

To avoid this ambiguity, we assume that there exists an order on the set of all domain variables we use. The field variables, too, are supposed to be taken from an ordered set. Then there will be only one translation of t where u and v appear sorted as arguments of f^l. If, for instance, $u < v$ in the given order, then t will be translated by $\mathsf{f}^l(u,v)$.

To prepare a formal treatment of the translation outlined here, we first fix the necessary additional conventions regarding the use of variable symbols. To obtain an order on the variables used, we simply restrict the translation to formulas only containing domain variables from the ordered set $\{v_1, v_2, \ldots\}$, and field variables from $\{x_1, x_2, \ldots\}$.

If $\boldsymbol{v}$ is any tuple of domain variables, then $\sigma\boldsymbol{v}$ denotes the permutation of $\boldsymbol{v}$ with sorted components. Analogously for tuples of field variables. For two tuples $\boldsymbol{v}$,$\boldsymbol{w}$, $\sigma(\boldsymbol{v}\cup\boldsymbol{w})$ is the sorted tuple whose components are the variables occuring either in $\boldsymbol{v}$ or in $\boldsymbol{w}$. Example: $\sigma((v_3,v_1),(v_1,v_5)) = (v_1,v_3,v_5)$.

In addition to these variables we will also use sets of auxiliary domain variables $\{\alpha_1,\alpha_2,\ldots\}$ and field variables $\{\zeta_1,\zeta_2,\ldots\}$. The only tuples of auxiliary domain variables that we are going to encounter will be of the special form $(\alpha_1,\alpha_2,\ldots,\alpha_k)$ for some k. For this reason we introduce the additional convention that $\boldsymbol{\alpha}^k$ denotes this specific tuple. The parameter k, i.e. the length of $\boldsymbol{\alpha}$, will usually be apparent from the context, in which case we simply write $\boldsymbol{\alpha}$. Similarly, by convention, $\boldsymbol{\zeta} = \boldsymbol{\zeta}^k = (\zeta_1,\ldots,\zeta_k)$ for the appropriate k.

The new vocabulary S_∞ is defined inductively. In each step new function and constant symbols are derived from the language obtained by the previous step. All the new function symbols will be of sort $\mathrm{D}^k\mathrm{F}^l$ for some $k \geq 0,\ l \geq 1$. To begin, define

$$\mathrm{S}_0 := \mathrm{S} \cup \{0,1,+,\cdot,\leq\}.$$

Each n-ary function and relation symbol of S is of sort D^{n+1} and D^n respectively, while $+$ and $\cdot$ are of sort F^3, $\leq$ is of sort F^2, and 0,1 are of sort F.

Now assume that S_n has been defined. For a formula $\phi\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle \in \mathrm{L}^{g}_{\mathrm{S}_n}$, the *pattern* of ϕ with respect to $\boldsymbol{v}$, written $\mathrm{p}(\phi;\boldsymbol{v})$, is the formula obtained by substituting the auxiliary variables $\boldsymbol{\alpha}$ for $\sigma\boldsymbol{v}$, and $\boldsymbol{\zeta}$ for $\sigma\boldsymbol{x}$ in ϕ:

$$\mathrm{p}(\phi;\boldsymbol{v}) :\equiv \phi[\sigma\boldsymbol{v}/\boldsymbol{\alpha}\,,\sigma\boldsymbol{x}/\boldsymbol{\zeta}].$$

Note that $\mathrm{p}(\phi;\boldsymbol{v})$ only depends on $\boldsymbol{v}$ as a set, not on their order in $\boldsymbol{v}$.

For each pattern $\mathrm{p}(\phi;v)$ of a formula $\phi\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle \in \mathrm{L}_{\mathrm{S}_n} \setminus \mathrm{L}_{\mathrm{S}_{n-1}}$ with $\boldsymbol{w} \neq \emptyset$ (define $\mathrm{L}_{\mathrm{S}_{-1}} = \emptyset$ for the first step of the induction), and each permutation $\pi\boldsymbol{w}$ of $\boldsymbol{w}$ let

$$\mathsf{f}^{\mathrm{p}(\phi;\boldsymbol{v}),\pi\boldsymbol{w}}$$

be a new function symbol of sort $\mathrm{D}^{|\boldsymbol{v}|}\ \mathrm{F}^{|\boldsymbol{x}|+1}$. If $\boldsymbol{v} = \boldsymbol{x} = \emptyset$, then $\mathsf{f}^{\mathrm{p},\boldsymbol{w}}$ is a constant symbol of sort F.

Let S_{n+1} denote the union of S_n and the set of these new symbols. Finally, define

$$\mathrm{S}_\infty := \bigcup_{n\in\mathbf{N}} \mathrm{S}_n \quad\text{and}\quad \mathrm{S}_+ := \mathrm{S}_\infty \setminus \mathrm{S}_0.$$

Example 2.5.4 The pattern of

$$\phi \equiv \mathsf{R}v_3v_2v_4v_1 \in \mathrm{L}_{\mathrm{S}_0}$$

with respect to v_2 and v_3 is

$$\mathrm{p}_0 \equiv \mathrm{p}(\phi;(v_3,v_2)) \equiv \mathsf{R}\alpha_2\alpha_1v_4v_1$$

Thus, the two new function symbols

$$\mathsf{f}^{\mathrm{p}_0,(v_4,v_3)} \quad\text{and}\quad \mathsf{f}^{\mathrm{p}_0,(v_3,v_4)}$$

are contained in S_1.

Next, a translation from terms and formulas t,ϕ of $\mathrm{L}^{g}_{\mathrm{S}}$ to terms and formulas t^*,ϕ^* of $\mathrm{L}_{\mathrm{S}_\infty}$ is defined:

Domain-terms: If t is a domain term, then $\mathrm{t}^* :\equiv \mathrm{t}$.
Atomic domain formulas: If ϕ is an atomic domain formula, then $\phi^* :\equiv \phi$.
Boolean operators: $(\phi\wedge\psi)^* :\equiv \phi^*\wedge\psi^*$ and $(\neg\phi)^* :\equiv \neg\phi^*$.
Quantification: $(\exists v\phi)^* :\equiv \exists v\phi^*$ and $(\exists x\phi)^* :\equiv \exists x\phi^*$.
Field-terms: **(a)** $x^* :\equiv x$.
(b) $0^* :\equiv 0 \quad 1^* :\equiv 1$
(c) $(\mathrm{t}_1+\mathrm{t}_2)^* :\equiv \mathrm{t}_1^*+\mathrm{t}_2^*, \quad (\mathrm{t}_1\cdot\mathrm{t}_2)^* :\equiv \mathrm{t}_1^*\cdot\mathrm{t}_2^*$.
(d) $([\phi\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle]_{\boldsymbol{w}})^* :\equiv \mathsf{f}^{\mathrm{p}(\phi^*;\boldsymbol{v}),\boldsymbol{w}}(\sigma\boldsymbol{v},\sigma\boldsymbol{x})$.
Atomic field formulas: $(\mathrm{t}_1\leq\mathrm{t}_2)^* :\equiv \mathrm{t}_1^*\leq\mathrm{t}_2^*$.

Observe that the translations t^* and ϕ^* have the same free variables as t and ϕ, respectively.

Example 2.5.5 Let S contain a three-placed relation symbol R, let

$$\phi :\equiv \forall x v_2 \left([\mathsf{R}v_3v_1v_2]_{(v_1,v_3)} \geq x \cdot [[\mathsf{R}v_3v_1v_2]_{v_1} \geq x]_{v_3}\right).$$

ϕ states that the Fubini-property holds for certain sets definable via R and is therefore a tautology of L^σ. From the formula

$$\phi_0\langle v_1, v_2, v_3\rangle :\equiv \mathsf{R}v_3v_1v_2 \ \in \mathrm{L}_{\mathrm{S}_0}$$

the two patterns

$$\begin{aligned} &\mathrm{p}_0 :\equiv \mathrm{p}(\phi_0; (v_2, v_3)) \equiv \mathsf{R}\alpha_2 v_1 \alpha_1 \quad \text{and} \\ &\mathrm{p}_1 :\equiv \mathrm{p}(\phi_0; v_2) \equiv \mathsf{R}v_3 v_1 \alpha_1 \end{aligned}$$

are derived, which give rise to function symbols $\mathsf{f}^{\mathrm{p}_0,v_1}$, $\mathsf{f}^{\mathrm{p}_1,(v_1,v_3)} \in \mathrm{S}_1$. Since $\phi_0^* \equiv \phi_0$, we get

$$\begin{aligned} &([\mathsf{R}v_3v_1v_2]_{v_1})^* \equiv \mathsf{f}^{\mathrm{p}_0,v_1}(v_2, v_3) \quad \text{and} \\ &([\mathsf{R}v_3v_1v_2]_{(v_1,v_3)})^* \equiv \mathsf{f}^{\mathrm{p}_1,(v_1,v_3)}(v_2). \end{aligned}$$

The pattern of

$$\phi_1\langle v_2, v_3, x\rangle :\equiv ([\mathsf{R}v_3v_1v_2]_{v_1} \geq x)^* \equiv \mathsf{f}^{\mathrm{p}_0,v_1}(v_2, v_3) \geq x \ \in \mathrm{L}_{\mathrm{S}_1}$$

with respect to v_2 is

$$\mathrm{p}_2 :\equiv \mathrm{p}(\phi_1; v_2) \equiv \mathsf{f}^{\mathrm{p}_0,v_1}(\alpha_1, v_3) \geq \zeta_1,$$

so that $\mathsf{f}^{\mathrm{p}_2,v_3}$ is an element of S_2 of sort DF^2. Then

$$([[\mathsf{R}v_3v_1v_2]_{v_1} \geq x]_{v_3})^* \equiv \mathsf{f}^{\mathrm{p}_2,v_3}(v_2, x),$$

and finally

$$\phi^* \equiv \forall x v_2 \left(\mathsf{f}^{\mathrm{p}_1,(v_1,v_3)}(v_2) \geq x \cdot \mathsf{f}^{\mathrm{p}_2,v_3}(v_2, x)\right).$$

We will also be needing an inverse mapping $(\cdot)^{-1}$ that translates S_∞- terms and formulas back into $\mathrm{L}^\sigma_\mathrm{S}$. Like $(\cdot)^*$, the inverse will be preserving free variables. For the definition of $(\cdot)^{-1}$ an order on the variables is not needed. Therefore ψ^{-1} will also be defined for $\psi \in \mathrm{L}_{\mathrm{S}_\infty}$ containing auxiliary variables $\alpha_i, \zeta_j, \ldots$; specifically the inverse p^{-1} of a pattern p becomes an $\mathrm{L}^\sigma_\mathrm{S}$-formula containing auxiliary variables $\alpha_i, \zeta_j, \ldots$.

We proceed by combining an induction on the number n for which $\mathrm{t}, \psi \in \mathrm{L}_{\mathrm{S}_n}$ with an induction on the structure of t and ψ.

The base case $n = 0$ is trivial. For t an S_0-term, ψ an S_0-formula, simply define $\mathrm{t}^{-1} :\equiv \mathrm{t}$, $\psi^{-1} :\equiv \psi$. Now assume that t^{-1} and ψ^{-1} have been defined for S_{n-1}- terms and formulas, and that this translation preserves free variables.

Specifically, if $\mathsf{f} = \mathsf{f}^{\mathrm{p},\boldsymbol{w}} \in \mathrm{S}_n$, then the pattern $\mathrm{p} = \mathrm{p}\langle \boldsymbol{\alpha}, \boldsymbol{w}, \boldsymbol{\zeta}\rangle$ is a formula in $\mathrm{L}_{\mathrm{S}_{n-1}}$ containing auxiliary variables $\boldsymbol{\alpha}$ and $\boldsymbol{\zeta}$. By induction hypothesis, $\mathrm{p}^{-1} \in \mathrm{L}^\sigma_\mathrm{S}$ is defined and contains just the free variables $\boldsymbol{\alpha}, \boldsymbol{w}$, and $\boldsymbol{\zeta}$.

Now let t be an S_n-term. The only nontrivial step in the inductive definition of t^{-1} is the case

$$\mathrm{t} \equiv \mathrm{t}(\boldsymbol{v}, \boldsymbol{x}) \equiv \mathsf{f}^{\mathrm{p},\boldsymbol{w}}(\mathrm{t}_1, \ldots, \mathrm{t}_k, \mathrm{s}_1, \ldots, \mathrm{s}_l)$$

with $\mathsf{f}^{\mathrm{p},\boldsymbol{w}} \in \mathrm{S}_n$ of sort $\mathrm{D}^k\mathrm{F}^{l+1}$, $\mathrm{t}_1(\boldsymbol{v}),\ldots,\mathrm{t}_k(\boldsymbol{v}) \in \mathrm{L}_{\mathrm{S}_n}$ terms of sort D, and $\mathrm{s}_1(\boldsymbol{v},\boldsymbol{x}),\ldots,\mathrm{s}_l(\boldsymbol{v},\boldsymbol{x}) \in \mathrm{L}_{\mathrm{S}_n}$ terms of sort F. The $\mathrm{t}_i(\boldsymbol{v})$ are in fact in $\mathrm{L}_{\mathrm{S}_0}$, and $(\mathrm{t}_i(\boldsymbol{v}))^{-1} = \mathrm{t}_i(\boldsymbol{v})$. By induction hypothesis (of the syntactic induction in $\mathrm{L}_{\mathrm{S}_n}$), $(\mathrm{s}_i(\boldsymbol{v},\boldsymbol{x}))^{-1} \in \mathrm{FT}^g_{\mathrm{S}}$ is defined for $i = 1,\ldots,l$. Let

$$\begin{aligned}\mathbf{t}^{-\mathbf{1}} &:= ((\mathrm{t}_1(\boldsymbol{v}))^{-1},\ldots,(\mathrm{t}_k(\boldsymbol{v}))^{-1}),\\ \mathbf{s}^{-\mathbf{1}} &:= ((\mathrm{s}_1(\boldsymbol{v},\boldsymbol{x}))^{-1},\ldots,(\mathrm{s}_l(\boldsymbol{v},\boldsymbol{x}))^{-1}).\end{aligned}$$

Define

$$(\mathsf{f}^{\mathrm{p},\boldsymbol{w}}(\mathrm{t}_1,\ldots,\mathrm{t}_k,\mathrm{s}_1,\ldots,\mathrm{s}_l))^{-1} :\equiv [\mathrm{p}^{-1}]_{\boldsymbol{w}}[\boldsymbol{\alpha}\,/\mathbf{t}^{-\mathbf{1}},\boldsymbol{\zeta}\,/\mathbf{s}^{-\mathbf{1}}].$$

Once the inverse t^{-1} is defined for S_n-terms, the inductive definition of ψ^{-1} for $\psi \in \mathrm{L}_{\mathrm{S}_n}$ is trivial, and effectively consists of replacing all field-terms in ψ by their inverse.

Example 2.5.6 Let

$$\psi(v_1,v_2) \equiv \mathsf{f}^{\mathrm{p}_0,v_1}(v_1,0.5,\mathsf{f}^{\mathrm{p}_0,v_1}(v_2,0.3,1)) < 1.$$

with

$$\mathrm{p}_0 \equiv \zeta_1 \leq \zeta_2 \wedge \mathsf{R}v_1\alpha_1 \in \mathrm{L}_{\mathrm{S}_0}.$$

Then $\mathrm{p}_0^{-1} \equiv \mathrm{p}_0$, and

$$(\mathsf{f}^{\mathrm{p}_0,v_1}(v_2,0.3,1))^{-1} \equiv [\mathrm{p}_0]_{v_1}[\alpha_1/v_2,\ \zeta_1/0.3,\ \zeta_2/1] \equiv [0.3 \leq 1 \wedge \mathsf{R}v_1v_2]_{v_1}.$$

Thus

$$\begin{aligned}\psi^{-1} &\equiv [\mathrm{p}_0]_{v_1}[\alpha_1/v_1,\ \zeta_1/0.5,\ \zeta_2/[0.3 \leq 1 \wedge \mathsf{R}v_1v_2]_{v_1}] < 1\\ &\equiv [0.5 \leq [0.3 \leq 1 \wedge \mathsf{R}v_1v_2]_{v_1} \wedge \mathsf{R}v_3v_1]_{v_3} < 1.\end{aligned}$$

When ψ^{-1} is translated back into $\mathrm{L}_{\mathrm{S}_\infty}$, the result obviously can not again be ψ, because $(\cdot)^*$ never produces f-terms ($\mathsf{f} \in \mathrm{S}_+$) with arguments other than variables.

To compute $(\psi^{-1})^*$, let

$$\mathrm{p}_1 \equiv 0.3 \leq 1 \wedge \mathsf{R}v_1\alpha_1.$$

Then

$$(0.5 \leq [0.3 \leq 1 \wedge \mathsf{R}v_1v_2]_{v_1} \wedge \mathsf{R}v_3v_1)^* \equiv 0.5 \leq \mathsf{f}^{\mathrm{p}_1,v_1}(v_2) \wedge \mathsf{R}v_3v_1.$$

With

$$\mathrm{p}_2 \equiv 0.5 \leq \mathsf{f}^{\mathrm{p}_1,v_1}(\alpha_2) \wedge \mathsf{R}v_3\alpha_1$$

we get

$$(\psi^{-1})^* \equiv \mathsf{f}^{\mathrm{p}_2,v_3}(v_1,v_2) < 1.$$

What about $((\psi^{-1})^*)^{-1}$? To find this formula, first the inverse of p_2 has to be computed. With $\mathrm{p}_1^{-1} \equiv \mathrm{p}_1$, this is

$$\begin{aligned}\mathrm{p}_2^{-1} &\equiv 0.5 \leq [\mathrm{p}_1]_{v_1}[\alpha_1/\alpha_2] \wedge \mathsf{R}v_3\alpha_1\\ &\equiv 0.5 \leq [0.3 \leq 1 \wedge \mathsf{R}v_1\alpha_2]_{v_1} \wedge \mathsf{R}v_3\alpha_1.\end{aligned}$$

Thus,

$$\begin{aligned}((\psi^{-1})^*)^{-1} &\equiv [\mathrm{p}_2^{-1}]_{v_3}[\alpha_1/v_1,\ \alpha_2/v_2] < 1\\ &\equiv [0.5 \leq [0.3 \leq 1 \wedge \mathsf{R}v_1v_2]_{v_1} \wedge \mathsf{R}v_3v_1]_{v_3} < 1\\ &\equiv \psi^{-1}.\end{aligned}$$

This example shows that $(\cdot)^{-1}$ is not one-one. The two different formulas ψ and $(\psi^{-1})^*$ are mapped to the same L_S^g- formula. When applied to a formula of the form ϕ^* with $\phi \in \mathrm{L}_\mathrm{S}^g$ it gave the original formula ϕ as a result. Lemma 2.5.8 shows that this always is the case, thereby giving a justification for calling $(\cdot)^{-1}$ the inverse (of $(\cdot)^*$).

The following technical lemma will be needed for the proof of lemma 2.5.10.

Lemma 2.5.7 Let $\phi\langle \boldsymbol{v},\boldsymbol{x}\rangle \in \mathrm{L}_{\mathrm{S}_\infty}$, $\boldsymbol{v}', \boldsymbol{x}'$ be tuples of domain and field variables. Then

$$\models_\sigma (\phi[\boldsymbol{v}/\boldsymbol{v}',\, \boldsymbol{x}/\boldsymbol{x}'])^{-1} \leftrightarrow \phi^{-1}[\boldsymbol{v}/\boldsymbol{v}',\, \boldsymbol{x}/\boldsymbol{x}'] \tag{2.23}$$

Proof: The proof is by induction on the structure of ϕ. The key step is to show that

$$\begin{aligned}\models_\sigma &(\mathsf{f}^{\mathrm{p},\boldsymbol{w}}(\mathrm{t}_1(\boldsymbol{v}),\ldots,\mathrm{t}_k(\boldsymbol{v}),\mathrm{s}_1(\boldsymbol{v},\boldsymbol{x}),\ldots,\mathrm{s}_l(\boldsymbol{v},\boldsymbol{x}))[\boldsymbol{v}/\boldsymbol{v}',\, \boldsymbol{x}/\boldsymbol{x}'])^{-1}\\ &= (\mathsf{f}^{\mathrm{p},\boldsymbol{w}}(\mathrm{t}_1(\boldsymbol{v}),\ldots,\mathrm{t}_k(\boldsymbol{v}),\mathrm{s}_1(\boldsymbol{v},\boldsymbol{x}),\ldots,\mathrm{s}_l(\boldsymbol{v},\boldsymbol{x})))^{-1}[\boldsymbol{v}/\boldsymbol{v}',\, \boldsymbol{x}/\boldsymbol{x}']\end{aligned} \tag{2.24}$$

for $\mathsf{f}^{\mathrm{p},\boldsymbol{w}} \in \mathrm{S}_+$. Using the notation

$$\mathbf{t}(\boldsymbol{v}) = (\mathrm{t}_1(\boldsymbol{v}),\ldots,\mathrm{t}_k(\boldsymbol{v})),\ \mathbf{s}(\boldsymbol{v},\boldsymbol{x}) = (\mathrm{s}_1(\boldsymbol{v},\boldsymbol{x}),\ldots,\mathrm{s}_l(\boldsymbol{v},\boldsymbol{x}))$$

the first term in 2.24 is

$$[\mathrm{p}^{-1}]_{\boldsymbol{w}}[\boldsymbol{\alpha}\,/\mathbf{t}[\boldsymbol{v}/\boldsymbol{v}'], \boldsymbol{\zeta}\,/\mathbf{s}[\boldsymbol{v}/\boldsymbol{v}',\, \boldsymbol{x}/\boldsymbol{x}']].$$

By lemma 2.5.3 this is logically equivalent to

$$[\mathrm{p}^{-1}]_{\boldsymbol{w}}[\boldsymbol{\alpha}\,/\mathbf{t}(\boldsymbol{v}),\, \boldsymbol{\zeta}\,/\mathbf{s}(\boldsymbol{v},\boldsymbol{x})][\boldsymbol{v}/\boldsymbol{v}',\, \boldsymbol{x}/\boldsymbol{x}'],$$

which is just the second term in (2.24). This proves (2.24). We omit the remaining steps of the induction. □

Lemma 2.5.8 For all $\phi \in \mathrm{L}_\mathrm{S}^g$: $(\phi^*)^{-1} \equiv \phi$.

Proof: The proof of the lemma basically relies on the fact that when $\psi \in \mathrm{L}_{\mathrm{S}_\infty}$ is of the form $\psi \equiv \phi^*$, all the substitutions performed in the definition of ψ^{-1} are trivial, i.e. do not require any renaming of bound variables.

For a formal proof, it is useful to prove a slightly stronger statement for ϕ together with an analogous statement about field terms $\mathrm{t} \in \mathrm{FT}_\mathrm{S}^g$.

(+) For all $\mathrm{t}(\boldsymbol{v},\boldsymbol{x}) \in \mathrm{FT}_\mathrm{S}^g$, tuples of domain variables $\boldsymbol{v}_{\mathrm{new}}$, and field variables $\boldsymbol{x}_{\mathrm{new}}$

$$(\mathrm{t}^*[\sigma\boldsymbol{v}/\boldsymbol{v}_{\mathrm{new}}, \sigma\boldsymbol{x}/\boldsymbol{x}_{\mathrm{new}}])^{-1} \equiv \mathrm{t}[\sigma\boldsymbol{v}/\boldsymbol{v}_{\mathrm{new}}, \sigma\boldsymbol{x}/\boldsymbol{x}_{\mathrm{new}}].$$

(++) For all $\phi(\boldsymbol{v},\boldsymbol{x}) \in \mathrm{L}_\mathrm{S}^g$, tuples of domain variables $\boldsymbol{v}_{\mathrm{new}}$, and field variables $\boldsymbol{x}_{\mathrm{new}}$, for which either $\boldsymbol{v}_{\mathrm{new}}(i) = \sigma\boldsymbol{v}(i)$, or $\boldsymbol{v}_{\mathrm{new}}(i)$ is a new variable not appearing in ϕ (analogously for $\boldsymbol{x}_{\mathrm{new}}$):

$$(\phi^*[\sigma\boldsymbol{v}/\boldsymbol{v}_{\mathrm{new}}, \sigma\boldsymbol{x}/\boldsymbol{x}_{\mathrm{new}}])^{-1} \equiv \phi[\sigma\boldsymbol{v}/\boldsymbol{v}_{\mathrm{new}}, \sigma\boldsymbol{x}/\boldsymbol{x}_{\mathrm{new}}].$$

The proof is by induction on the structure of t and ϕ. The expression $[\sigma\boldsymbol{v}/\boldsymbol{v}_{\text{new}}, \sigma\boldsymbol{x}/\boldsymbol{x}_{\text{new}}]$ henceforward is abbreviated by $[\ldots]$.

(a): Let t, ϕ be first-order. Since both $(\cdot)^*$ and $(\cdot)^{-1}$ are the identity in this case, (+) and (++) trivially hold.

(b): Let $\mathrm{t} \in \mathrm{FT}^{\sigma}_{\mathrm{S}}$. If $\mathrm{t} \equiv x$ then (+) is clearly true. For $\mathrm{t} \equiv \mathrm{t}_1 + \mathrm{t}_2$ we have

$$\begin{aligned}
(\mathrm{t}^*[\ldots])^{-1} &\equiv ((\mathrm{t}_1^* + \mathrm{t}_2^*)[\ldots])^{-1} \\
&\equiv (\mathrm{t}_1^*[\ldots] + \mathrm{t}_2^*[\ldots])^{-1} \\
&\equiv (\mathrm{t}_1^*[\ldots]^{-1} + \mathrm{t}_2^*[\ldots]^{-1}) \\
&\equiv \mathrm{t}_1[\ldots] + \mathrm{t}_2[\ldots] && \text{(by induction hypothesis)} \\
&\equiv (\mathrm{t}_1 + \mathrm{t}_2)[\ldots] \\
&\equiv \mathrm{t}[\ldots]
\end{aligned}$$

The case $\mathrm{t} \equiv \mathrm{t}_1 \cdot \mathrm{t}_2$ is treated analogously. Now let t be of the form $[\phi(\boldsymbol{v},\boldsymbol{w},\boldsymbol{x})]_{\boldsymbol{w}}$. Then

$$\begin{aligned}
(\mathrm{t}^*[\ldots])^{-1} &\equiv (\mathsf{f}^{\mathrm{p}(\phi^*;\boldsymbol{v}),\boldsymbol{w}}(\sigma\boldsymbol{v}, \sigma\boldsymbol{x})[\ldots])^{-1} \\
&\equiv (\mathsf{f}^{\mathrm{p}(\phi^*;\boldsymbol{v}),\boldsymbol{w}}(\boldsymbol{v}_{\text{new}}, \boldsymbol{x}_{\text{new}}))^{-1} \\
&\equiv [\mathrm{p}(\phi^*;\boldsymbol{v})^{-1}]_{\boldsymbol{w}}\,[\boldsymbol{\alpha}\,/\boldsymbol{v}_{\text{new}}, \boldsymbol{\zeta}\,/\boldsymbol{x}_{\text{new}}] \\
&\equiv [(\phi^*[\sigma\boldsymbol{v}/\boldsymbol{\alpha}\,, \sigma\boldsymbol{x}/\boldsymbol{\zeta}\,])^{-1}]_{\boldsymbol{w}}\,[\boldsymbol{\alpha}\,/\boldsymbol{v}_{\text{new}}, \boldsymbol{\zeta}\,/\boldsymbol{x}_{\text{new}}] \\
&\equiv [\phi[\sigma\boldsymbol{v}/\boldsymbol{\alpha}\,, \sigma\boldsymbol{x}/\boldsymbol{\zeta}\,]]_{\boldsymbol{w}}\,[\boldsymbol{\alpha}\,/\boldsymbol{v}_{\text{new}}, \boldsymbol{\zeta}\,/\boldsymbol{x}_{\text{new}}] && \text{(ind.hypoth. for } \phi) \\
&\equiv [\phi]_{\boldsymbol{w}}\,[\sigma\boldsymbol{v}/\boldsymbol{\alpha}\,, \sigma\boldsymbol{x}/\boldsymbol{\zeta}\,][\boldsymbol{\alpha}\,/\boldsymbol{v}_{\text{new}}, \boldsymbol{\zeta}\,/\boldsymbol{x}_{\text{new}}] && (\boldsymbol{w} \cap \boldsymbol{\alpha}\, = \emptyset) \\
&\equiv \mathrm{t}[\ldots] && \text{(by lemma 2.5.3)}
\end{aligned}$$

(c): Let $\phi \in \mathrm{L}^{\sigma}_{\mathrm{S}}$. If ϕ is an atomic domain formula, then (++) trivially holds. When $\phi \equiv \mathrm{t}_1 \leq \mathrm{t}_2$ is an atomic field formula, we get

$$\begin{aligned}
(\phi^*[\ldots])^{-1} &\equiv ((\mathrm{t}_1^* \leq \mathrm{t}_2^*)[\ldots])^{-1} \\
&\equiv (\mathrm{t}_1^*[\ldots] \leq \mathrm{t}_2^*[\ldots])^{-1} \\
&\equiv \mathrm{t}_1[\ldots] \leq \mathrm{t}_2[\ldots] && \text{(ind.hypoth. for } \mathrm{t}_1, \mathrm{t}_2) \\
&\equiv (\mathrm{t}_1 \leq \mathrm{t}_2)[\ldots] \\
&\equiv \phi[\ldots]
\end{aligned}$$

The induction steps for $\phi \equiv \phi_1 \wedge \phi_2$ and $\phi \equiv \neg\psi$ are trivial. For $\phi \equiv \exists u \psi(\boldsymbol{v}, u, \boldsymbol{x})$ we get

$$\begin{aligned}
(\phi^*[\ldots])^{-1} &\equiv ((\exists u \psi^*)[\ldots])^{-1} \\
&\equiv (\exists u(\psi^*[\ldots]))^{-1} && (u \notin \boldsymbol{v}_{\text{new}}) \\
&\equiv \exists u(\psi^*[\ldots])^{-1} \\
&\equiv \exists u(\psi[\ldots]) && \text{(ind.hypoth. for } \psi) \\
&\equiv (\exists u \psi)[\ldots] && (u \notin \boldsymbol{v}_{\text{new}}) \\
&\equiv \phi[\ldots]
\end{aligned}$$

Identically for quantification over field variables.

□

2.5.3 Corresponding Structures

After the purely syntactical considerations of the previous section, we now consider the correspondence of statistical S-structures and a certain class of S_∞-structures.

First we give a canonical construction of a two sorted S_∞-structure

$$\mathfrak{M}^* = (M^*, F^*, I^*)$$

from a statistical S-structure $\mathfrak{M}= (M, I, \mathfrak{F}, (\mathfrak{A}_n, \mu_n)_n)$.

The basic idea is very simple: we leave the standard parts (M, I) and $\mathfrak{F}$ of $\mathfrak{M}$ unchanged, and substitute suitable interpretations of the new symbols S_+ for the measure component $(\mathfrak{A}_n, \mu_n)_n$ of $\mathfrak{M}$.

Thus, let $(M^*, I^* \restriction S) := (M, I)$ and $(F^*, I^* \restriction \{0, 1, +, \cdot, \leq\}) := \mathfrak{F}$.

It remains to define the interpretations of function symbols in S_+. If $\mathsf{f} = \mathsf{f}^{\mathrm{p}, \boldsymbol{w}}$ is of sort $\mathrm{D}^k\mathrm{F}^{l+1}$, then p^{-1} is an $\mathrm{L}^\sigma_\mathrm{S}$-formula with free variables $\boldsymbol{\alpha}$, $\boldsymbol{w}$, and $\boldsymbol{\zeta}$. Thus, we may define for $\boldsymbol{a} \in \mathrm{M}^k$, $\boldsymbol{r} \in \mathrm{F}^l$:

$$I^*(\mathsf{f}) : \; (\boldsymbol{a}, \boldsymbol{r}) \mapsto (\mathfrak{M}, \boldsymbol{\alpha} / \boldsymbol{a}, \boldsymbol{\zeta} / \boldsymbol{r})([\mathrm{p}^{-1}]_{\boldsymbol{w}}).$$

The following lemma demonstrates the close correspondence between $\mathfrak{M}$ and $\mathfrak{M}^*$.

Lemma 2.5.9 Let $\mathfrak{M}$ be a statistical S-structure, $\mathfrak{M}^*$ the corresponding S_∞-structure. Then

(i) for each S_∞-term $\mathrm{t}(\boldsymbol{v},\boldsymbol{x})$:

$$(\mathfrak{M}^*, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\mathrm{t}) = (\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\mathrm{t}^{-1}),$$

(ii) for each S_∞-formula $\phi(\boldsymbol{v},\boldsymbol{x})$:

$$\mathfrak{M}^*(\phi(\boldsymbol{v},\boldsymbol{x})) = \mathfrak{M}(\phi^{-1}(\boldsymbol{v}, \boldsymbol{x})),$$

(iii) for each S_∞- sentence ϕ:

$$\mathfrak{M}^* \models_\sigma \phi \;\text{ iff }\; \mathfrak{M} \models \phi^{-1}.$$

Proof: (i): By induction on the structure of t. The only nontrivial step is

$$\mathrm{t} \equiv \mathsf{f}^{\mathrm{p}, \boldsymbol{w}}(\mathrm{t}_1(\boldsymbol{v}), \ldots, \mathrm{t}_k(\boldsymbol{v}), \mathrm{s}_1(\boldsymbol{v}, \boldsymbol{x}), \ldots, \mathrm{s}_l(\boldsymbol{v}, \boldsymbol{x})).$$

Assume that (i) holds for the $\mathrm{t}_i, \mathrm{s}_j$. Abbreviating $(\mathfrak{M}^{(*)}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})$ by $(\mathfrak{M}^{(*)} \ldots)$, then

$$\begin{aligned}
(\mathfrak{M}^* \ldots)(\mathrm{t}) &= I^*(\mathsf{f}^{\mathrm{p}, \boldsymbol{w}})((\mathfrak{M}^* \ldots)(\mathrm{t}_1), \ldots, (\mathfrak{M}^* \ldots)(\mathrm{s}_l)) \\
&= I^*(\mathsf{f}^{\mathrm{p}, \boldsymbol{w}})((\mathfrak{M} \ldots)(\mathrm{t}_1^{-1}), \ldots, (\mathfrak{M} \ldots)(\mathrm{s}_l^{-1})) \\
&= \left(\mathfrak{M}, \alpha_1/(\mathfrak{M} \ldots)(\mathrm{t}_1^{-1}), \ldots, \zeta_l/(\mathfrak{M} \ldots)(\mathrm{s}_l^{-1})\right)([\mathrm{p}^{-1}]_{\boldsymbol{w}}) \\
&= (\mathfrak{M} \ldots)([\mathrm{p}^{-1}]_{\boldsymbol{w}}[\alpha_1/\mathrm{t}_1^{-1}, \ldots, \zeta_l/\mathrm{s}_l^{-1}]) \\
&= (\mathfrak{M} \ldots)(\mathrm{t}^{-1})
\end{aligned}$$

(ii) and (iii) immediately follow from (i) by induction on the structure of ϕ. □

Lemma 2.5.9 (iii) is the first important building block for the reduction of L^σ. The second one will be a first-order axiomatization of the class of S_∞-structures that correspond to statistical S-structures. More precisely, we shall define a recursively enumerable set of axioms $\mathrm{AX} \subset \mathrm{L}_{\mathrm{S}_\infty}$, such that for every S_∞-structure $\mathfrak{M}$: if $\mathfrak{M} \models \mathrm{AX}$ then a statistical S-structure $\mathfrak{M}^{-1}$ can be defined so that an analogy of lemma 2.5.9 holds for $\mathfrak{M}$ and $\mathfrak{M}^{-1}$.

The axioms in AX are divided into four groups. In the first group are the axioms of real closed fields:

- Let RCF be as in definition 1.2.3.

The second group of axioms make sure that the interpretations of the $\mathsf{f} \in \mathrm{S}_+$ can be used to induce a function on definable subsets of the domain. We call this collection of axioms SF for "Set Function".

- For all $\phi\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle, \psi\langle \boldsymbol{v}', \boldsymbol{w}', \boldsymbol{x}'\rangle \in \mathrm{L}_{\mathrm{S}_\infty}$ with $|\,\boldsymbol{w}\,| = |\,\boldsymbol{w}'\,|$ let SF contain the axiom

$$\forall \boldsymbol{v}\boldsymbol{v}'\boldsymbol{x}\boldsymbol{x}' \left(\forall \boldsymbol{u}(\phi[\boldsymbol{w}/\boldsymbol{u}] \leftrightarrow \psi[\boldsymbol{w}'/\boldsymbol{u}]) \rightarrow \\ \mathsf{f}^{\mathrm{p}(\phi;\boldsymbol{v}),\boldsymbol{w}}(\sigma\boldsymbol{v}, \sigma\boldsymbol{x}) = \mathsf{f}^{\mathrm{p}(\psi;\boldsymbol{v}'),\boldsymbol{w}'}(\sigma\boldsymbol{v}', \sigma\boldsymbol{x}')\right) \tag{2.25}$$

  where $\boldsymbol{u}$ is a tuple of variables with $\boldsymbol{u} \cap \boldsymbol{v} = \boldsymbol{u} \cap \boldsymbol{v}' = \emptyset$.

That the $\mathsf{f} \in \mathrm{S}_+$ define not only a set function, but a probability measure is ensured by the axioms PM:

- Let PM contain
  - For all $\mathsf{f} \in \mathrm{S}_+$ the axiom
$$\forall \boldsymbol{v}\boldsymbol{x}\, \mathsf{f}(\boldsymbol{v}, \boldsymbol{x}) \geq 0. \tag{2.26}$$
  - For each n and $\tau_n :\equiv (w_1 = w_1 \wedge \ldots \wedge w_n = w_n)$ the axiom
$$\mathsf{f}^{\mathrm{p}(\tau_n;\emptyset),\boldsymbol{w}} = 1. \tag{2.27}$$
  - For all $\phi\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle, \psi\langle \boldsymbol{v}', \boldsymbol{w}, \boldsymbol{x}'\rangle \in \mathrm{L}_{\mathrm{S}_\infty}$ the axiom
$$\forall \boldsymbol{v}\boldsymbol{v}'\boldsymbol{x}\boldsymbol{x}' \Big(\neg\exists \boldsymbol{w}(\phi \wedge \psi) \rightarrow \mathsf{f}^{\mathrm{p}(\phi\vee\psi;\boldsymbol{v}\cup\boldsymbol{v}'),\boldsymbol{w}}(\sigma(\boldsymbol{v} \cup \boldsymbol{v}'), \sigma(\boldsymbol{x} \cup \boldsymbol{x}')) \\ = \mathsf{f}^{\mathrm{p}(\phi;\boldsymbol{v}),\boldsymbol{w}}(\sigma\boldsymbol{v}, \sigma\boldsymbol{x}) + \mathsf{f}^{\mathrm{p}(\psi;\boldsymbol{v}'),\boldsymbol{w}}(\sigma\boldsymbol{v}', \sigma\boldsymbol{x}')\Big) \tag{2.28}$$

The next set of axioms, HOM, provides for the homogeneity of the probability measure:

- For every pattern p and every pair of function symbols $\mathsf{f}^{\mathrm{p},\pi_1\boldsymbol{w}}$, $\mathsf{f}^{\mathrm{p},\pi_2\boldsymbol{w}}$, HOM contains the axiom
$$\forall \boldsymbol{v}\boldsymbol{x}\, (\mathsf{f}^{\mathrm{p},\pi_1\boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{x}) = \mathsf{f}^{\mathrm{p},\pi_2\boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{x}))\,. \tag{2.29}$$

Finally, the Fubini-property must be axiomatized.

- For $\phi\langle \boldsymbol{v}, \boldsymbol{w}', \boldsymbol{w}'', \boldsymbol{x}\rangle \in \mathrm{L}_{\mathrm{S}_\infty}$ let

$$\phi'\langle \boldsymbol{v}, \boldsymbol{w}', \boldsymbol{x}, y\rangle :\equiv \mathsf{f}^{\mathrm{p}(\phi;\boldsymbol{v},\boldsymbol{w}'),\boldsymbol{w}''}(\sigma(\boldsymbol{v}\cup\boldsymbol{w}'), \sigma\boldsymbol{x}) \geq y.$$

  where $y \notin \boldsymbol{x}$. FUB then contains all the axioms of the form

$$\forall \boldsymbol{v}\boldsymbol{x}y \left(\mathsf{f}^{\mathrm{p}(\phi;\boldsymbol{v}),(\boldsymbol{w}',\boldsymbol{w}'')}(\sigma\boldsymbol{v}, \sigma\boldsymbol{x}) \leq y \cdot \mathsf{f}^{\mathrm{p}(\phi';\boldsymbol{v}),\boldsymbol{w}'}(\sigma\boldsymbol{v}, \sigma(\boldsymbol{x}\cup y))\right). \tag{2.30}$$

Finally, let AX := RCF ∪ SF ∪ PM ∪ HOM ∪ FUB. Clearly, AX is a recursively enumerable set of axioms.

Lemma 2.5.10 Let $\mathfrak{M}$ be a statistical S-structure. Then $\mathfrak{M}^* \models$ AX.

Proof: For the proof compute the inverses of the given axioms and check that they are tautologies in L^σ. The assertion then follows by lemma 2.5.9. We only give the details for the axioms of FUB.

Let ψ be an instance of the FUB-schema given by the formula $\phi\langle \boldsymbol{v}, \boldsymbol{w}', \boldsymbol{w}'', \boldsymbol{x}\rangle$. To simplify notation, assume that $\boldsymbol{v} = (v_0, \ldots, v_{k-1})$, $\boldsymbol{w}' = (v_k, \ldots, v_{k+l-1})$, $\boldsymbol{w}'' = (v_{k+l}, \ldots, v_{k+l+m-1})$, and $\boldsymbol{x} = (x_0, \ldots, x_{n-1})$. Then

$$\begin{aligned}
(\mathsf{f}^{\mathrm{p}(\phi;\boldsymbol{v}),(\boldsymbol{w}',\boldsymbol{w}'')}(\boldsymbol{v}, \boldsymbol{x}))^{-1} &\equiv [\mathrm{p}(\phi;\boldsymbol{v})^{-1}]_{(\boldsymbol{w}',\boldsymbol{w}'')}[\boldsymbol{\alpha}^k/\boldsymbol{v}, \boldsymbol{\zeta}^n/\boldsymbol{x}] \\
(\mathsf{f}^{\mathrm{p}(\phi';\boldsymbol{v}),\boldsymbol{w}'}(\boldsymbol{v}, \boldsymbol{x}, y))^{-1} &\equiv [(\mathrm{p}(\phi';\boldsymbol{v}))^{-1}]_{\boldsymbol{w}'}[\boldsymbol{\alpha}^k/\boldsymbol{v}, \boldsymbol{\zeta}^{n+1}/(\boldsymbol{x}, y)] \\
\mathrm{p}(\phi';\boldsymbol{v})^{-1} &\equiv \left(\mathsf{f}^{\mathrm{p}(\phi;(\boldsymbol{v},\boldsymbol{w}')),\boldsymbol{w}''}(\boldsymbol{v}, \boldsymbol{w}', \boldsymbol{x}) \geq y\ [\boldsymbol{v}/\boldsymbol{\alpha}^k, (\boldsymbol{x}, y)/\boldsymbol{\zeta}^{n+1}]\right)^{-1} \\
&\equiv [\mathrm{p}(\phi;(\boldsymbol{v},\boldsymbol{w}'))^{-1}]_{\boldsymbol{w}''}[(\alpha_k, \ldots, \alpha_{k+l-1})/\boldsymbol{w}'] \geq \zeta_{n+1}.
\end{aligned}$$

Putting things together,

$$\begin{aligned}
\psi^{-1} \equiv \forall \boldsymbol{v}\boldsymbol{x}y \Big(&[(\mathrm{p}(\phi;\boldsymbol{v}))^{-1}]_{(\boldsymbol{w}',\boldsymbol{w}'')} \geq \\
& y \cdot [[\mathrm{p}(\phi;(\boldsymbol{v},\boldsymbol{w}'))^{-1}]_{\boldsymbol{w}''}[(\alpha_k, \ldots, \alpha_{k+l-1})/\boldsymbol{w}'] \geq y]_{\boldsymbol{w}'}\Big) [\boldsymbol{\alpha}^k/\boldsymbol{v}, \boldsymbol{\zeta}^n/\boldsymbol{x}].
\end{aligned}$$

By lemma 2.5.7

$$\begin{aligned}
&\models_\sigma \mathrm{p}(\phi;\boldsymbol{v})^{-1} \leftrightarrow \phi^{-1}[\boldsymbol{v}/\boldsymbol{\alpha}^k, \boldsymbol{x}/\boldsymbol{\zeta}^n] \\
&\models_\sigma \mathrm{p}(\phi;(\boldsymbol{v},\boldsymbol{w}'))^{-1} \leftrightarrow \phi^{-1}[(\boldsymbol{v},\boldsymbol{w}')/\boldsymbol{\alpha}^{k+1}, \boldsymbol{x}/\boldsymbol{\zeta}^n],
\end{aligned}$$

so that ψ^{-1} is logically equivalent to

$$\begin{aligned}
\forall \boldsymbol{v}\boldsymbol{x}y \Big(\Big(&[\phi^{-1}[\boldsymbol{v}/\boldsymbol{\alpha}^k, \boldsymbol{x}/\boldsymbol{\zeta}^n]]_{(\boldsymbol{w}',\boldsymbol{w}'')} \geq \\
& y \cdot [[\phi^{-1}[(\boldsymbol{v},\boldsymbol{w}')/\boldsymbol{\alpha}^{k+1}, \boldsymbol{x}/\boldsymbol{\zeta}^n]]_{\boldsymbol{w}''}[(\alpha_k, \ldots, \alpha_{k+l-1})/\boldsymbol{w}'] \geq y]_{\boldsymbol{w}'}\Big) [\boldsymbol{\alpha}^k/\boldsymbol{v}, \boldsymbol{\zeta}^n/\boldsymbol{x}]\Big) \\
&\equiv \forall \boldsymbol{v}\boldsymbol{x}y \left([\phi^{-1}]_{(\boldsymbol{w}',\boldsymbol{w}'')} \geq y \cdot [[\phi^{-1}]_{\boldsymbol{w}''} \geq y]_{\boldsymbol{w}'}\right)
\end{aligned}$$

The last formula clearly is a tautology of L^σ. □

Now assume that $\mathfrak{M} = (M, F, I)$ is an S_∞-structure with $\mathfrak{M} \models$ AX. We construct a corresponding statistical S-structure $\mathfrak{M}^{-1} = (M^{-1}, I^{-1}, \mathfrak{F}, (\mathfrak{A}_n, \mu_n)_n)$.

The standard part of $\mathfrak{M}^{-1}$ is straightforward:

$$(M^{-1}, I^{-1}) := (M, I\restriction S) \quad \text{and} \quad \mathfrak{F} := (F, I\restriction\{0, 1, +, \cdot, \leq\}).$$

The measure component $(\mathfrak{A}_n, \mu_n)_n$ must, of course, be gained from the interpretations of S_+ in $\mathfrak{M}$.

Let $\mathfrak{A}_n$ consist of all subsets of M^n that are first-order definable by an S_∞-formula with parameters:

$$\mathfrak{A}_n := \{(\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\phi\langle \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{x}\rangle) \mid \phi\langle \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{x}\rangle \in \mathrm{L}_{S_\infty}, \boldsymbol{a} \in M^{|\boldsymbol{v}|}, \boldsymbol{r} \in F^{|\boldsymbol{x}|}\}.$$

It is easy to see that $\mathfrak{A}_n$ is an algebra, because of the correspondence between set theoretic intersection and complementation on the one hand, and syntactic conjunction and negation on the other.

For $A = (\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\phi\langle \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{x}\rangle) \in \mathfrak{A}_n$ define

$$\mu_n(A) := (\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\mathsf{f}^{\mathrm{p}(\phi;\boldsymbol{v}),\boldsymbol{w}}(\sigma\boldsymbol{v}, \sigma\boldsymbol{x})). \tag{2.31}$$

Since $\mathfrak{M} \models$ SF, μ_n is well defined, i.e. $\mu_n(A)$ does not depend on the particular choice of ϕ. From $\mathfrak{M} \models$ PM it follows easily that μ_n is a probability measure on $\mathfrak{A}_n$. Also, it is immediate from $\mathfrak{M} \models$ HOM that μ_n satisfies homogeneity.

$(\mathfrak{A}_n)_n$ is closed under products because the product of

$$\begin{aligned} A &= (\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\phi\langle \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{x}\rangle) \quad \text{and} \\ B &= (\mathfrak{M}, \boldsymbol{v}'/\boldsymbol{a}', \boldsymbol{x}'/\boldsymbol{r}')(\phi'\langle \boldsymbol{v}', \boldsymbol{w}', \boldsymbol{x}'\rangle) \end{aligned}$$

where, without loss of generality, $(\boldsymbol{v} \cup \boldsymbol{w}) \cap (\boldsymbol{v}' \cup \boldsymbol{w}') = \boldsymbol{x} \cap \boldsymbol{x}' = \emptyset$, can be defined by $(\phi \wedge \phi')\langle \boldsymbol{v}, \boldsymbol{v}', \boldsymbol{w}, \boldsymbol{w}', \boldsymbol{x}, \boldsymbol{x}'\rangle$.

It remains to show that $(\mu_n)_n$ has the Fubini property, which we shall do in a little more detail.

Let A=$(\mathfrak{M}$,$\boldsymbol{v}/\boldsymbol{a}$,$\boldsymbol{x}/\boldsymbol{s})$ $(\phi\langle \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{x}\rangle) \in \mathfrak{A}_n$, $I \subset \{1, \ldots, n\}$ with $1 \leq k := \mid I \mid < n$ and $\boldsymbol{a}' \in M^I$. It must be shown that $\sigma^I_{\boldsymbol{a}'}(A) \in \mathfrak{A}_{n-k}$, $A_{I,r} \in \mathfrak{A}_k$ for all $r \in [0, 1]$, and $\mu_n(A) \geq r\,\mu_k(A_{\mathrm{I},r})$.

Since the $\mathfrak{A}_n$ are closed under permutations, and μ_n is homogeneous for every n, it suffices to consider the case $I = \{1, \ldots, k\}$. With

$$\boldsymbol{w}' := (\boldsymbol{w}(1), \ldots, \boldsymbol{w}(k)) \quad \text{and} \quad \boldsymbol{w}'' := (\boldsymbol{w}(k+1), \ldots, \boldsymbol{w}(n))$$

we then have

$$\sigma^I_{\boldsymbol{a}'}(\mathrm{A}) = (\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{w}'/\boldsymbol{a}', \boldsymbol{x}/\boldsymbol{s})\langle\phi(\boldsymbol{v}, \boldsymbol{w}', \boldsymbol{w}'', \boldsymbol{x}\rangle) \in \mathfrak{A}_{n-k}.$$

As in the definition of FUB define

$$\phi'\langle \boldsymbol{v}, \boldsymbol{w}', \boldsymbol{x}, y\rangle \equiv \mathsf{f}^{\mathrm{p}(\phi;\boldsymbol{v},\boldsymbol{w}'),\boldsymbol{w}''}(\sigma(\boldsymbol{v}, \boldsymbol{w}'), \sigma\boldsymbol{x}) \geq y.$$

Then

$$\begin{aligned} A_{I,r} &= \{\boldsymbol{a}' \in M^I \mid \mu_{n-k}(\sigma^I_{\boldsymbol{a}'}(A)) \geq r\} \\ &= \{\boldsymbol{a}' \in M^I \mid (\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{w}'/\boldsymbol{a}', \boldsymbol{x}/\boldsymbol{s}) \left(\mathsf{f}^{\mathrm{p}(\phi;(\boldsymbol{v},\boldsymbol{w}')),\boldsymbol{w}''}(\sigma(\boldsymbol{v} \cup \boldsymbol{w}'), \sigma\boldsymbol{x})\right) \geq r\} \\ &= \{\boldsymbol{a}' \in M^I \mid (\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{w}'/\boldsymbol{a}', \boldsymbol{x}/\boldsymbol{s}, y/r) \left(\mathsf{f}^{\mathrm{p}(\phi;(\boldsymbol{v},\boldsymbol{w}')),\boldsymbol{w}''}(\sigma(\boldsymbol{v} \cup \boldsymbol{w}'), \sigma\boldsymbol{x}) \geq y\right)\} \\ &= (\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{s}, y/r)(\phi'(\boldsymbol{v}, \boldsymbol{w}', \boldsymbol{x}, y)) \\ &\in \mathfrak{A}_k. \end{aligned}$$

By definition

$$\begin{aligned} \mu_k(A_{I,r}) &= (\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{s}, y/r) \left(\mathsf{f}^{\mathrm{p}(\phi';\boldsymbol{v}),\boldsymbol{w}'}(\sigma\boldsymbol{v}, \sigma(\boldsymbol{x}, y))\right) \quad \text{and} \\ \mu_n(A) &= (\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{s})(\mathsf{f}^{\mathrm{p}(\phi;\boldsymbol{v}),\boldsymbol{w}}(\sigma\boldsymbol{v}, \sigma\boldsymbol{x})) \end{aligned}$$

Since $\mathfrak{M} \models$ FUB, we have

$$(\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{s}, y/r) \models \mathsf{f}^{\mathrm{p}(\phi;\boldsymbol{v}),\boldsymbol{w}}(\sigma\boldsymbol{v}, \sigma\boldsymbol{x}) \geq y \cdot \mathsf{f}^{\mathrm{p}(\phi';\boldsymbol{v}),\boldsymbol{w}'}(\sigma\boldsymbol{v}, \sigma(\boldsymbol{x} \cup y)),$$

and thus

$$\mu_n(\mathrm{A}) \geq r \cdot \mu_k(\mathrm{A}_{I,r}).$$

We now have shown that $(\mathfrak{A}_n, \mu_n)_n$ is a sequence of algebras and measures that satisfies the consistency conditions. This is not quite enough to prove that

$$\mathfrak{M}^{-1} := (M^{-1}, I^{-1}, \mathfrak{F}, (\mathfrak{A}_n, \mu_n)_n)$$

is a statistical S-structure: it remains to be shown that for any variable assignment γ and every field term $\mathrm{t} \in \mathrm{FT}^\sigma_{\mathrm{S}}$, $(\mathfrak{M}, \gamma)(\mathrm{t})$ is defined. Part (i) of the following lemma asserts that this is the case.

Lemma 2.5.11 Let $\mathfrak{M}$ be an S_∞-structure with $\mathfrak{M} \models$ AX.

(i) For all $\mathrm{t}(\boldsymbol{v},\boldsymbol{x}) \in \mathrm{FT}^\sigma_{\mathrm{S}}$, $\boldsymbol{a} \in \mathrm{M}^{|\boldsymbol{v}|}$, $\boldsymbol{r} \in \mathrm{F}^{|\boldsymbol{x}|}$: $(\mathfrak{M}^{-1}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\mathrm{t})$ is defined and equal to $(\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\mathrm{t}^*)$.

(ii) For all $\phi\langle\boldsymbol{v},\boldsymbol{x}\rangle \in \mathrm{L}^\sigma_{\mathrm{S}}$: $\mathfrak{M}^{-1}(\phi\langle\boldsymbol{v},\boldsymbol{x}\rangle) = \mathfrak{M}(\phi^*\langle\boldsymbol{v}, \boldsymbol{x}\rangle)$.

(iii) For all sentences ϕ in $\mathrm{L}^\sigma_{\mathrm{S}}$: $\mathfrak{M}^{-1} \models_\sigma \phi$ iff $\mathfrak{M} \models \phi^*$.

Proof: (i) and (ii) are proved simultaneously by induction on the structure of t and ϕ.

(a): For first-order t and ϕ (i) and (ii) are trivially true.

(b): Let $\mathrm{t}=[\phi\langle\boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle]_{\boldsymbol{w}}$. Then, by induction hypothesis, $(\mathfrak{M}^{-1}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\phi\langle\boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle)$ is definable in $\mathfrak{M}$ via ϕ^* and therefore in $\mathfrak{A}_{|\boldsymbol{w}|}$. Also, by definition

$$\begin{aligned} (\mathfrak{M}^{-1}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\mathrm{t}) &= \mu_{|\boldsymbol{w}|}((\mathfrak{M}^{-1}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\phi\langle\boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle)) \\ &= \mu_{|\boldsymbol{w}|}((\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\phi^*\langle\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{x}\rangle)) \\ &= (\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\mathsf{f}^{\mathrm{p}(\phi^*;\boldsymbol{v}),\boldsymbol{w}}(\sigma\boldsymbol{v}, \sigma\boldsymbol{x})) \\ &= (\mathfrak{M}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\mathrm{t}^*) \end{aligned}$$

(c): Let $\phi\langle \boldsymbol{v},\boldsymbol{x}\rangle \in \mathrm{L}_{\mathrm{S}}^{\sigma}$ and assume that (i) holds for every probability term appearing in ϕ. Then (ii) holds for every atomic subexpression of ϕ. The remaining induction steps on the structure of ϕ are trivial, so we conclude that (ii) holds for ϕ.

Part (iii) of the lemma is immediate from (ii). □

With the preceding lemma we are finally ready to formulate the reduction of $\mathscr{L}^{\sigma}$ to first-order logic in the following theorem.

Theorem 2.5.12 For all $\Phi \subset \mathrm{L}_{\mathrm{S}}^{\sigma}$, $\phi \in \mathrm{L}_{\mathrm{S}}^{\sigma}$:

$$\Phi \models_{\sigma} \phi \quad \text{iff} \quad \Phi^* \cup \mathrm{AX} \models \phi^*.$$

Proof: "⇒:" Let $\Phi \models_{\sigma} \phi$, $\mathfrak{M}$ be an S_{∞}-structure with $\mathfrak{M} \models \Phi^* \cup \mathrm{AX}$. Since $\mathfrak{M} \models \mathrm{AX}$, there is a corresponding statistical S-structure $\mathfrak{M}^{-1}$ with $\mathfrak{M}^{-1} \models_{\sigma} \Phi$ by lemma 2.5.11. Then, $\mathfrak{M}^{-1} \models_{\sigma} \phi$, and, again by lemma 2.5.11, $\mathfrak{M} \models \phi^*$.
"⇐:" Let $\Phi^* \cup \mathrm{AX} \models \phi^*$, $\mathfrak{M}$ a statistical S-structure with $\mathfrak{M} \models_{\sigma} \Phi$. By lemma 2.5.8 $\Phi = (\Phi^*)^{-1}$, so that with lemma 2.5.9 $\mathfrak{M}^* \models \Phi^*$. With $\mathfrak{M}^* \models \mathrm{AX}$ we conclude that $\mathfrak{M}^* \models \phi^*$, and once more by lemmas 2.5.8 and 2.5.9, $\mathfrak{M} \models_{\sigma} \phi$. □

What consequence does theorem 2.5.12 have for our understanding of $\mathscr{L}^{\sigma}$? Putting it severely, it may be said that L^{σ} basically is nothing but a shorthand notation for a certain class of first-order theories – a notation that certainly provides a tremendous improvement in the readability of formulas.

The main benefit we gain from the reduction of $\mathscr{L}^{\sigma}$ is its inheritance of many nice properties of first-order logic. First of all, we get a completeness result for $\mathscr{L}^{\sigma}$: A complete formal proof-system for first-order logic, together with the translation rules for $(\cdot)^*$ and $(\cdot)^{-1}$, and the new set of axioms AX, provides a complete formal proof-system for $\mathscr{L}^{\sigma}$. Furthermore, the two most important consequences of the completeness proof for first-order logic – compactness and the Löwenheim - Skolem - theorem – are seen to also hold for $\mathscr{L}^{\sigma}$. (For the Löwenheim-Skolem-theorem observe that the transformation $\mathfrak{M} \to \mathfrak{M}^*$ preserves the domain and hence the cardinality.)

The proof-system for $\mathscr{L}^{\sigma}$ that has been outlined above, which includes translating back and forth between $\mathrm{L}_{\mathrm{S}}^{\sigma}$ and $\mathrm{L}_{\mathrm{S}_{\infty}}$, would certainly be very awkward to work with. However, we can do without these translations by using a proof-system for $\mathscr{L}^{\sigma}$ that consists of a proof-system for first-order logic handling probability terms $[\phi]_{\boldsymbol{w}}$ just like ordinary terms (except for the extra explanation for how to perform substitution into such a term), and the translated set of axioms AX^{-1}. The resulting proof system will then look the same as the system given in [Bacchus, 1990a]. Bacchus proved completeness for his system directly. In the light of the results of this section it is not surprising that his completeness proof is essentially a standard completeness proof for first-order logic, with the necessary extra considerations added for probability terms.

Note that none of the material contained in this section depends on the fact that we assume probabilities to take values in real closed fields. The only important point here is that they take values in a definable class of structures, and that they are only required to be finitely additive. Therefore, the same completeness result obtains if we were to use probabilities in a simpler type of structures than rc-fields, which can be interesting either for the reason that a simpler "qualitative" notion of probability is intended to be formalized, or in order to obtain simpler inference procedures for automated deduction.

Chapter 3

Default Reasoning About Probabilities: An Analysis

3.1 Subjective Probabilities and Degrees of Belief

For statistical probabilities it has been fairly easy to give a precise definition of their meaning. The semantics of probabilistic statements that refer to what we have decided to call subjective probabilities is much more difficult to define.

The following three examples certainly fall into the category of subjective probability statements.

> "The probability that a one, two, or three has turned up at this throw of the die is 0.3"
> "The probability that this film is an American production is ≤ 0.5."
> "With a high probability his father was an actor."

The manifest difference between these statements and the statistical probability statements considered in section 2.1 is that here there is no mentioning of a general class of events in which a relative frequency of a specific property could be observed. Rather, the subject of each of our new statements is one specific event – a specific toss of a certain die, an individual film we are currently watching, a certain person (we continue to use the more abstract term "event" to designate any kind of subject in a subjective probability statement, be it an event like the toss of a die, or an object like a mystery film). For each of these events it is either true or false that it has the property under consideration. The probabilities stated describe the uncertainty of our knowledge, or our *degree of belief*, about which of these two alternatives holds in reality.

Different people will usually have different information pertaining to a given event, so that their judgment of the likelihood that it possesses a certain property may be quite different. This is in clear contrast to the statistical probabilities: they are understood to describe objective features of the world, and any two people who make a statement of their value should, in principle, agree as to their magnitude.

This alone does not completely justify the use of the term subjective probability because one may argue that a personal degree of belief – at least by a perfect rational agent – can be derived as a necessary consequence from the given partial knowledge according to some objective principles of inductive reasoning (as sought by Carnap [1950], for instance). However, since experience tells us that even two sufficiently rational people sharing the same evidence may arrive at substantially different probability evaluations, and, furthermore, generally neither of the two persons will be able to "prove" to the other one that his or her probability evaluation is the "true" one, it seems inevitable to acknowledge a certain amount of subjectivity when a non-0,1 truth value is attached to a proposition.

All this brings us only marginally closer to a definition of the meaning of the term probability as used in the examples above. As yet, we only have argued that it is descriptive of some subjective degree of belief in the truth of a proposition. It is apparent, though, that attempting to define subjective probability simply by reducing it to terms like degree of belief, degree of confirmation, or judgment of likelihood is inadequate, because such a definition of subjective probability would use terms our intuitive understanding of which is no better than of the one we started with, and in whose definitions, in turn, the term probability is likely to be used. Before we approach a non-circular definition of the two terms "degree of belief" and " subjective probability", it is useful to first have a closer look at how these are related.

Comparing the two expressions "degree of belief" and "subjective probability", a trivial, yet important, difference is observed: the latter certainly designates a probability, i.e. an additive function with values in [0,1]; it is not immediately clear, however, that this must also be true for the former.

An argument in favour of assuming degrees of belief to also be probabilities, and in fact to treat the terms degree of belief and subjective probability as synonyms, can be derived from subjectivistic interpretations of probability pioneered by Ramsey (e.g.[1931]) and de Finetti (e.g.[1937]). De Finetti is most closely associated with the interpretation of (subjective) probabilities in terms of *betting odds*: If an individual is offered two bets, one of which offers a gain of G for a stake $p \cdot \mathrm{G}$ $(0 \leq p \leq 1)$ in case of an event E coming true, the other one offering the same gain G for a stake $(1-p) \cdot \mathrm{G}$ for the case that E does not occur, then clearly the first bet is preferable for $p = 0$, and the second for $p = 1$. For exactly one $p \in [0, 1]$ the person offered the bet will be indifferent towards choosing the one or the other bet. That value of p then is defined as his degree of belief in E.

The first conclusion we can draw from this betting odds scenario is that there is a way to always associate a precise numerical value with a degree of belief. No matter how great an individual's uncertainty about the event E is, there is a way to extort from him a statement about his beliefs that allows us to describe it by a single numerical value.

Furthermore, it can be shown (see [de Finetti, 1937]) that when the individual's beliefs are *coherent*, i.e. do not make him accept a series of bets in which he is certain to lose, then his degrees of belief must obey the rules of probability calculus (this has come to be known as the "Dutch Book Theorem").

With such an operational definition of degrees of belief as outlined here, i.e. a definition in terms of a procedure for how to measure beliefs, there is the danger that the numbers obtained do not only represent a person's epistemic state, but also depend on the method by which they are determined. This is quite obvious in the betting odds - paradigm: the preference for a bet on E or not E may not only depend on our assessment of the likelihood of E, but also on our willingness to risk the larger of the two stakes $p \cdot \mathrm{G}$ and $(1-p) \cdot \mathrm{G}$. Ramsey writes that a degree of belief "has no precise meaning unless we specify more exactly how it is to be measured"([Ramsey, 1931]). He then proceeds to develop an alternative system for measuring beliefs that avoids the betting odds method's problem of unequal utilities.

From the classical subjectivistic interpretations of probability we may therefore conclude: there are ways to equate a person's degrees of belief with a unique subjective probability measure, and thus to view the two terms as synonymous. However, such identifications require specific methods by which degrees of belief are measured, and the resulting subjective probability measure is not guaranteed to only depend on an agent's epistemic state, but may also be influenced by the method employed.

An important criticism brought against attempts to identify degrees of belief with a single (subjective) probability measure is that these must necessarily fail to distinguish between degrees of belief founded on knowledge, and degrees of belief based on ignorance (e.g.[Shafer, 1976]).

For an example of a degree of belief based on knowledge, suppose that a person is asked about the likelihood that a coin we just flipped did turn up heads. We may expect a rather confident and matter-of-fact answer that this probability is 0.5. Also, if we had not put an outright question for the assessment of this likelihood, but had extracted this degree of belief by one or another of the methods mentioned above, then we may expect to always receive the same result.

Compare this scenario to the following: a person with none but the most rudimentary knowledge of botany is presented with the facts that every angiosperm is either a monocotyledon, or a dicotyledon, and that bellis perennis is an angiosperm. Asked about the likelihood for bellis perennis being a monocotyledon, that person would probably feel incapable of making any statement. Resorting to one of the methods for forcing our subject to make a statement that we can translate into a numerical value for his degree of belief for bellis perennis being an angiosperm, we may again obtain the value 0.5 (but not necessarily so). This degree of belief, being only a product of the complete ignorance of the concepts involved, clearly has a rather different nature than the one assigned to the outcome of the coin-flip.

The level of ignorance that accompanies a person's evaluation for the likelihood of an event to have a certain property is not reflected in a model of degrees of belief as a single probability measure. This problem is one of the motivations for the development of *Dempster-Shafer theory* ([Shafer, 1976]): here an epistemic state is represented by a *belief function Bel* which is interpreted to measure the amount of positive confirmation obtained for a proposition. Missing confirmation for a proposition ϕ is not assumed to be confirmation for its negation, so that usually $Bel(\phi)+Bel(\neg\phi)<1$. With this formalism, the agent's beliefs in the first of the two examples above may be modeled by a belief function with $Bel(\mathsf{Heads})=Bel(\mathsf{Tails})=1/2$, while the epistemic state of the agent in the second example is best represented by $Bel(\mathsf{Monoc})=Bel(\mathsf{Dic})=0$.

It has been observed early on ([Dempster, 1967],[Fagin and Halpern, 1991]) that such belief functions correspond to certain sets of probability measures: the value of $Bel(\phi)$ can be seen as the lower bound of $\{\nu(\phi) \mid \nu \in P\}$ with P a set of probability measures. Thus, in Dempster-

Shafer theory, degrees of belief are essentially modeled by certain sets of probability measures.

This leaves us with already two different explanations for the relationship between degrees of belief and subjective probability: either the two terms can be treated as synonymous, or a degree of belief can be understood as being comprised of several subjective probability measures. Formalizations in which degrees of belief are no longer represented by numerical values (e.g. [Darwiche and Ginsberg, 1992]) draw up an even more tenuous connection between the two terms.

3.2 Interpreting Degrees of Belief by Thought Experiments

We will now proceed to develop an interpretation of subjective probabilities and degrees of belief rather different from the traditional ones mentioned above. Instead of concentrating on the manifestations of degrees of belief as preferences between choices, we put at the core of this interpretation a process by which an agent may arrive at a degree of belief in the first place.

Also, our interpretation will be more cautious than the classical ones in that degrees of belief will not be identified with single probability values. Instead, similarly as in Dempster-Shafer theory, the picture of degrees of belief as defined by a set of subjective probability measures will emerge from the interpretation here presented. With this understanding of the relationship between degrees of belief and subjective probability, the second example statement (p. 60) is seen to express a degree of belief *of* [0,0.5] for the film under consideration being American, while the subjective probability for this being true only is partially specified to lie *in* [0,0.5].

It might be added, that the proposed view of subjective probabilities has a rather frequentistic flavour, and by that token alone, follows an intuition differing from classical subjectivism. For us, introducing the concept of relative frequencies into our understanding of subjective probabilities is instrumental for linking the formation of an individual's degrees of belief to his or her statistical knowledge.

The following two examples will serve as a guide towards an epistemological interpretation of degrees of belief and default reasoning about probabilities.

Example 3.2.1 I'm playing a game of dice with a friend who just has made the roll of the die that will decide the game: if she rolls a four or better, she wins; if a three or less turns up, I win. The die has come to rest out of my sight, but the outcome has been observed by my friend. By the somewhat satisfied expression on her face I gather that I will less likely have won than lost this game. "Less likely" I'm here willing to quantify by a probability of 0.3, so that my degree of belief in the current toss t having the property $\rho :\equiv$ " Either a one, a two, or a three has turned up" is expressed by stating for it the probability 0.3.

Example 3.2.2 Scanning channels on TV we tune in to a mystery film. We just catch the last part of a spectacular car chase, apparently taking place in a European city. These observations induce us to state that the film has been an expensive production with probability ≥ 0.7, and is of American origin only with probability ≤ 0.5.

The two uncertain events described in these examples are of a somewhat different nature: the first one is a product of a what can only be understood as a random process. The uncertain event in the second example, however, is not random in the classical sense that a toss of a die or the drawing of a card from a shuffled deck is random. The film, a fragment of which we have happened to see on TV, is not broadcast at that time as a result of being drawn from a gigantic urn containing all mystery films. Given that we are partially ignorant of the deterministic chain of events that led to the screening of that particular film at that particular time, however, for us endows this events with all the features of randomness. Some partial knowledge we may possess of the actual chain of events causing the given observation, and the ignorance about some other of its parts, combines to the imperfect perception of that chain of events as a *random mechanism*.

Thus, both the tossing of a die and the selection of a film for screening by an anonymous producer according to rules unknown to us are instances of random mechanisms. The random mechanism associated with a certain event is determined by what information we have about the actual chain of events, particularly information that is relevant for predicting the chain of event's likely results. Our model of the random mechanism resulting in the broadcast of the observed mystery film would be altered, for instance, if we knew that the network screening the film was known for its close ties to Hollywood, or else that it had a reputation for showing lesser known European productions. In the die example, on the other hand, there does not seem to be any additional relevant information we might obtain about the toss of the die that would help us to construct a more accurate model of a random mechanism than that of an arbitrary die-toss; even complete knowledge about the initial parameters (position, speed, spin) of the particular toss observed will hardly enable us to devise an improved model.

An assignment of degrees of belief as given in our two examples now can be understood as being based on two elements: the model of a random mechanism associated with the observed event, and the specific evidence provided by the observation.

Figure 3.1 depicts the complete evidence presented to us in the scenario of example 3.2.1: the state of the game and the expression on our opponents face who knows the outcome of the roll of the die. Our assessment of the likelihood for that outcome to be three or less does not only depend on this evidence, but also on the assumption that the toss of the die was fair, i.e. on our concept of the underlying random mechanism. If we suspected that the die was heavily loaded in favour of turning up sixes, or that our opponent was able (and willing) to manipulate the outcome of the toss by throwing the die in a particular skilful way, then our model of the random mechanism would be altered, and our estimate for the likelihood of $\rho(\mathrm{t})$ somewhat smaller.

Analogously, besides the evaluation of the evidence provided by the glimpse we caught of one scene in the film, a model of a certain random mechanism underlies the assignment of degrees of belief in the film-example: a model basically determined by our complete ignorance of what the actual reasons were for broadcasting that particular film at that particular time. Clearly, the same evidence would induce us to assign different degrees of belief if we knew that the network screening the film currently was running a series "Low budget: how the Europeans make the best of it" – information that would change the model of the random mechanism.

The random mechanisms that have been described in the discussion of examples 3.2.1 and 3.2.2 were supposed to be perfectly well-defined in the sense that the unknown chain of

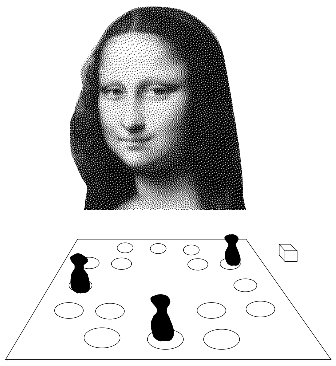

Figure 3.1: The evidence in example 3.2.1

events was reduced to the random draw from a set of well-defined alternatives (here the six possible outcomes of a die-toss and the set of all mystery films) according to well defined chances (here equal chances for all numbers on the die, and a chance for any film to be drawn determined by its likelihood to be shown on TV at an arbitrary time). It is not claimed, though, that this will always be the case. Our image of the random mechanism may well contain unknown parameters.

For an illustration of this, consider a variation of example 3.2.1: suppose that I have a vague suspicion that my friend has a loaded die up her sleeve that she occasionally will use instead of the fair one, and which enables her to roll a six in 90% of throws, a five, four, three, or two each in 2.2% of throws, and a one in 1.2% of throws. Finding myself in the same situation as described previously, I will now have to incorporate into my model of the random mechanism that produced the crucial toss of the die the possibility that in that toss the loaded die was in fact supplanted for the fair one. The result might be a model of a random mechanism consisting of first a random draw of one of either a fair or a loaded die, and a subsequent toss of that die. However, feeling unable to evaluate the likelihood for my friend to have cheated at the observed toss, I am unable to specify the respective chances for the two dice to be drawn. This makes my thought experiment depend on an unknown parameter ranging from 0 (the loaded die is certain to be drawn) to 1 (the fair die is certain to be drawn).

Another possibility to cope with the uncertainty of what kind of die has been used in the actual toss, is to stay with the model of a single draw from the six possible outcomes of die-tosses, but to leave unspecified according to which two different sets of chances outcomes are drawn: either according to a chance of 1/6 for each outcome, or according to a chance of 0.9 for a six, of 0.022 for five, four, three, and two, and 0.012 for one.

As yet, it has only been tried to isolate two components that the formation of degrees of belief about an individual event e depends on: the model of a random mechanism which produced e, and the observed properties of e. Putting the two parts together, gives rise to an interpretation of the meaning of degrees of belief.

Based on our model of a random mechanism, we can consider a long (hypothetical) sequence of events that are independent realizations of the same random mechanism. Moreover, we can imagine all the elements of the sequence to provide us with the same evidence as e. Such an imaginary sequence of events we call a *thought experiment.*

A property ϕ that e may or may not possess will occur with a certain relative frequency in this sequence. This relative frequency we call the subjective probability for e having ϕ. An agent's degree of belief for e having ϕ is the most specific prediction the agent is willing to make for the relative frequency of ϕ in the sequence.

Figure 3.2 illustrates a thought experiment for the specification of degrees of belief in example 3.2.1. In a long sequence of situations similar to the present one in that they are brought about by a toss of a fair die in the same state of the game and inspiring the same expression on the face of our opponent, we expect that in three out of ten situations the number cast will actually turn out to be one, two, or three, while in seven out of ten situations, four or more turns up.

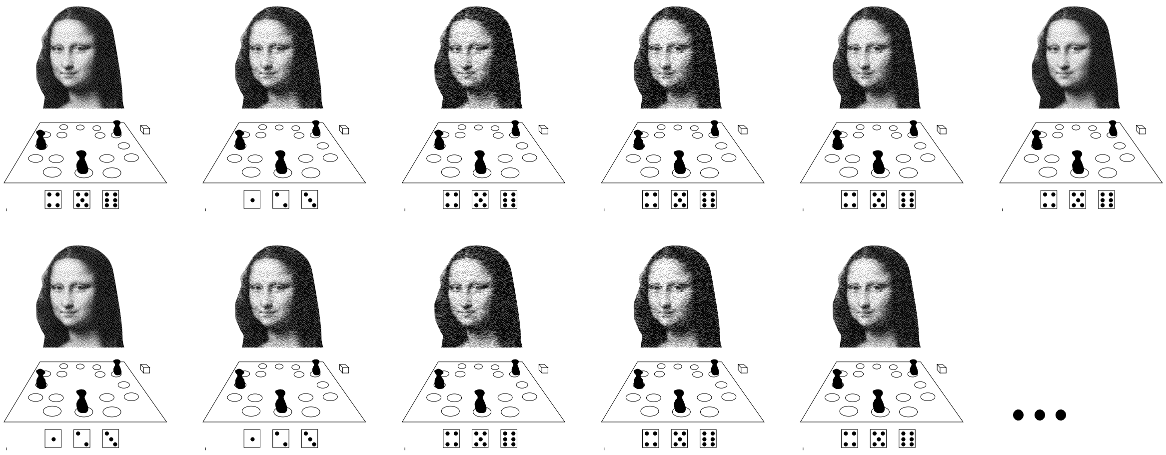

Figure 3.2: The thought experiment in example 3.2.1

For the film example, the thought experiment may consist of an imaginary sequence of film scenes accidentally seen on TV and showing a spectacular car chase in a rather European-looking environment. The degrees of belief then are defined by the bounds [0,0.5] and [0.7,1] we predict for the proportion of, respectively, American and expensive productions in this sequence.

Note that there is no inherent difference in these two examples that would cause degrees of belief to be point-valued in the first example, and interval-valued in the second. This distinction is merely brought about by the gratuitous assumptions that the agent in the first example for some reason feels confident on the basis of the available evidence to state an exact degree of belief, i.e. to make a very specific prediction for the outcome of the thought experiment, while the agent in the second example only feels able to give a partial description of that outcome.

Multi-valued degrees of belief will usually result when the parameters of the random mechanism are not fully known. Reconsider the modification of example 3.2.1 introduced above.

Suppose that under the assumption that a loaded die has been tossed, I would assign a probability of only 0.01 to the current toss having property ρ. In other words: if for the random mechanism associated with the current toss a random draw from the six possible outcomes according to the chances 0.9, 0.022, ..., 0.022, 0.012 is used, then the predicted frequency of one, two, or three in a series of realizations of this random process providing the described evidence is 0.01. Now consider the first of the two random mechanisms described for incorporating the uncertainty about which die has been used. Assuming that the unknown parameter in this random mechanism actually has the value 0 (i.e. the loaded die is assumed to have been cast) makes it equivalent to the fully determined random mechanism just described, and thus would entail a predicted frequency of 0.01 for ρ in the thought experiment. The complementary assumption of the value 1 for this parameter results in the old prediction 0.3 for this frequency. Values of the parameter in between 0 and 1 will continuously change the predicted frequency from 0.01 to 0.3, so that, in the absence of any assumptions for the parameter, the frequency can only be predicted to have some value in [0.01,0.3], which then is the degree of belief assigned to e having property ρ.

The result is somewhat different when the second model of a partially defined random mechanism mentioned above is used. Here there are only two distinct values for the sets of parameters that determine the random mechanism. One of the two values entails a predicted frequency of 0.01, the other of 0.3. Hence, from this model of a random mechanism we derive a degree of belief of $\{0.01, 0.3\}$ for e having ρ.

While at first it seems somewhat unusual to allow a degree of belief consisting of two isolated values, this is really quite desirable when it is our aim to supply semantics for a wide range of subjective belief statements. The statement "the probability that by this toss of the die I have won the game is either 0.01 or 0.3", for example, makes perfect sense in the given context. The speaker here just distinguishes two different hypothesis about how the given event has been produced, entailing two different probabilities.

The thought experiment interpretation of degrees of belief here developed is summarized in the following postulate.

Postulate 1: *The degree of belief that an uncertain event* e *has property* ϕ *is the predicted bound on the relative frequency of* ϕ *in a long (imaginary) sequence of events, each of which is an independent realization of the random mechanism modeling the chain of events that produced* e*, and each of which has the same properties that have been observed in* e.

In the formulation of this postulate, as in the discussion above, the somewhat imprecise, but intuitive, concept of a "long" sequence of events has been used. For a really strict definition of a degree of belief we would rather have to speak about the limiting behaviour of the relative frequency when the thought experiment is imagined to continue indefinitely. We will continue to speak somewhat sloppily of the relative frequency, or the proportion, of some property in a long sequence, when, what in fact is meant is the limit this frequency attains as the sequence length tends towards infinity.

Postulate 1 defines the meaning of a degree of belief. Alternatively, we might have chosen to define the meaning of subjective probability by the following modification of postulate 1.

Postulate 1b: *The subjective probability for an uncertain event* e *to have property ϕ is the relative frequency of ϕ in a long (imaginary) sequence of events, each of which is an independent realization of the random mechanism modeling the chain of events that produced* e*, and each of which has the same properties that have been observed in* e.

Observe the delicate distinction: on the one hand, by the agent there is perceived to be a unique subjective probability distribution determined by the outcome of the thought experiment. But even though the random mechanism that sets up the thought experiment is a creation of the agent's mind, and the thought experiment is performed mentally, it is not assumed that he or she will be able to make a precise prediction of its outcome! In other words: the thought experiment, despite of being an artefact of an agent's reasoning process, by the agent is perceived to have objective properties not completely known to the agent.

The thought experiment interpretation of the meaning of degrees of belief may be argued to be more versatile than a betting-odds interpretation in that it allows us to provide consistent semantics for a wider range of statements about degrees of belief. For an illustration of this, reconsider the statement $\phi :\equiv$ "the probability that by this toss of the die I have won the game is either 0.01 or 0.3". With the betting-odds interpretation, this statement, taken at its face value, will be understood to mean that I would consider either 0.01 or 0.3 as fair odds for a bet on my having won the game. However, on further interrogation, it would probably turn out that the odds o for which I will be indifferent towards betting on either the win or the loss of the current game are somewhere in between 0.01 and 03, so that the original statement would appear to have been patently false, a correct statement being $\phi' :\equiv$ "the probability that by this toss of the die I have won the game is o".

With the thought experiment interpretation, in contrast, both ϕ and ϕ' are equally coherent belief statements that can be understood as being inspired by different conceptions of the random mechanism.

Interpreting subjective probabilities in terms of relative frequencies, naturally, is an idea that appeals to holders of the frequency view of probability (provided they are willing to credit a subjective probability statement with any meaning at all). Reichenbach [1949], when he introduces the direct inference principle (cf. p.3), in fact, does not only propose it as a means to derive subjective probabilities, but as an explication of the meaning of (subjective) probability: the subjective probability that e has property ϕ is the statistical probability of ϕ in a (suitably chosen) reference class for e. By this interpretation, a statement of a subjective probability presupposes some well-defined statistical knowledge. Carnap [1950] basically supports this point of view, but demands that the subjective probability only is an estimate for the actual statistical probability, which may be unknown.

This interpretation obviously is closely related to the one given here. When the evidence provided by an uncertain event just establishes its belonging to a certain reference class, and the circumstances under which the event has been observed do not suggest anything different than a random draw from that reference class, then the thought experiment will just consist of an imagined actual drawing of samples from that reference class. A thought experiment of this kind is an instance of what Shafer and Tversky [1985] have called a "mental experiment": the mental performing of an experiment that, in principle, could be carried out in practice, and the use of its anticipated result for obtaining probability judgments.

Thought experiments, as developed here as a general explanation for subjective probabilities, also extend to cases where a subjective probability can no longer be interpreted as an (estimated) objective statistical probability. As an example consider an observation of a film f in which actors A and B appear together. Suppose that we know that 80% of the films in which A appears are American productions, but that this is only true for 20% of the films with actor B. Suppose, too, that we know that f actually is the only film in which A and B appear together. With this information we would probably assign to the film f being American a subjective probability somewhere in between 0.2 and 0.8, say 0.5. This can not be understood as the statistical probability in any existing reference class. However, in a thought experiment we are free to resort to considering a sequence of merely hypothetical films that all have actors A and B in them, but may differ with regard to their origin.

When applied to situations as this, the thought experiment - model certainly can only serve as a semantic interpretation of the meaning of a degree of belief, but will hardly provide a rule for their (numerical) computation. Particularly, the question of how to construct a random mechanism for the thought experiment, and how to translate the evidence into a predicted bias for the outcome of realizations of the random mechanism, are outside the scope of the interpretation given in postulate 1. On the other hand, while from that postulate little can be learned about how exactly evidence should be evaluated to determine subjective probabilities, it will turn out that from this interpretation of degrees of belief an explicit rule can be derived for how to utilize statistical background information in this task.

3.3 The Role of Statistical Information

Example 3.2.1 (continued): What, in the situation described previously, will be my degree of belief in the proposition $\rho_i(\mathrm{t}) :\equiv$ "i has turned up in the current toss t" $(i = 1, 2, 3)$? The observation I have made only provides evidence that bears on the probability of $\rho(\mathrm{t})$, but does not give me any reason to think one of the three alternatives $\rho_1(\mathrm{t}), \rho_2(\mathrm{t}), \rho_3(\mathrm{t})$ more likely than another. However, I do have the information that the statistical probability of each of the ρ_i in tosses of a fair die is 1/6. Specifically, this means that each of the ρ_i has an equal statistical probability. This statistical knowledge determines my prediction of the outcome of the thought experiment associated with the present event: I will expect that here, too, each of the three alternatives ρ_1, ρ_2, ρ_3 will appear with equal frequency $0.3/3 = 0.1$. Similarly, for $i = 4, 5, 6$, a degree of belief $0.7/3$ will be assigned to $\rho_i(\mathrm{t})$.

Example 3.2.2 (continued): While a commercial break has stopped the flow of useful information, we have time to make up our mind whether we want to continue watching that mystery film. Having a preference for films with a happy end, we first attempt to estimate the likelihood for this film to have one. None of the evidence provided in the short scene we have seen directly suggests either a happy or an unhappy ending. Fortunately, however, we do have recourse to statistical information with what relative frequency happy endings have occurred in the great number of mystery films (distinguished by their having combinations of the properties A ($\approx$ American) and E ($\approx$ expensive)) shown on television in the last few years. Using our syntax for the representation of statistical probabilities, let this information consist of

$$\begin{array}{cc} [\mathsf{HE}v \mid \neg\mathsf{A}v \wedge \neg\mathsf{E}v]_v = 0.5 & [\mathsf{HE}v \mid \neg\mathsf{A}v \wedge \mathsf{E}v]_v = 0.7 \\ [\mathsf{HE}v \mid \mathsf{A}v \wedge \neg\mathsf{E}v]_v = 0.7 & [\mathsf{HE}v \mid \mathsf{A}v \wedge \mathsf{E}v]_v = 0.9. \end{array} \tag{3.1}$$

Here it is no longer obvious what prediction for the relative frequency of happy endings in the thought experiment we should derive from these statistics and our prior predictions of the frequencies of A and E. It is easy, though, to obtain some bounds for the plausible values of this frequency.

For an upper bound we may suppose that in the hypothetical sequence of mystery films the relative frequency of those of the four properties $\neg\mathsf{A} \wedge \neg\mathsf{E}, \ldots, \mathsf{A} \wedge \mathsf{E}$ is maximal (within the given bounds that the relative frequency of property A is at most 0.5, and that of E at least 0.7) for which the conditional statistical probability $[\mathsf{HE}v \mid \ldots]_v$ has the greatest values. This is achieved by assuming an outcome of the thought experiment in which both the relative frequency of $\mathsf{A} \wedge \mathsf{E}$ and $\neg\mathsf{A} \wedge \mathsf{E}$ are 0.5, i.e. every film in fact turns out to be expensive, and the number of American films is maximal. For such a sequence then a relative frequency

$$[\mathsf{HE}v \mid \mathsf{A}v \wedge \mathsf{E}v]_v \cdot 0.5 + [\mathsf{HE}v \mid \neg\mathsf{A}v \wedge \mathsf{E}v]_v \cdot 0.5 = 0.45 + 0.35 = 0.8 \tag{3.2}$$

of happy endings should be predicted.

Similarly, by considering an outcome of the thought experiment in which the number of expensive or American films is minimal, i.e. in which the relative frequency of $\neg\mathsf{A} \wedge \mathsf{E}$ is 0.7, and that of $\neg\mathsf{A} \wedge \neg\mathsf{E}$ is 0.3, a lower bound of

$$[\mathsf{HE}v \mid \neg\mathsf{A}v \wedge \mathsf{E}v]_v \cdot 0.7 + [\mathsf{HE}v \mid \neg\mathsf{A}v \wedge \neg\mathsf{E}v]_v \cdot 0.3 = 0.49 + 0.15 = 0.64 \tag{3.3}$$

is obtained for the expected frequency of HE.

A plausible degree of belief for the film to have a happy end therefore is the interval [0.64,0.8].

None of the two extreme outcomes of the thought experiment here considered can be identified as unrealistic on the basis of the given statistical data; after all (3.1) allows us to consistently assume that both $\mathsf{A} \wedge \mathsf{E}$ and $\neg\mathsf{A} \wedge \mathsf{E}$ have a statistical probability of 0.5 (making (3.2) a very reasonable estimate), or that $\neg\mathsf{A} \wedge \mathsf{E}$ and $\neg\mathsf{A} \wedge \neg\mathsf{E}$ have statistical probabilities of 0.7 and 0.3, respectively (making (3.3) a very reasonable estimate). Thus, the statistical data will not be seen to warrant the specification of a degree of belief for $\mathsf{HE}\mathrm{f}$ more specific than [0.64,0.8].

If, on the other hand, further statistical information enables us to predict an outcome of the thought experiment in which the relative frequencies of $\neg\mathsf{A} \wedge \neg\mathsf{E}, \ldots, \mathsf{A} \wedge \mathsf{E}$ are more precisely specified, then a tighter bounded degree of belief for $\mathsf{HE}\mathrm{f}$ will be obtained. Assume, for instance, that it is also known that

$$[\mathsf{A}v \mid \mathsf{E}v]_v = 0.9 \quad [\mathsf{A}v]_v = 0.7 \quad [\mathsf{E}v]_v = 0.2. \tag{3.4}$$

(3.1) and (3.4) together define a unique statistical probability distribution on the properties HE, A, and E given by

$$\begin{array}{cccc} \neg\mathsf{HE}\wedge\neg\mathsf{A}\wedge\neg\mathsf{E} & \neg\mathsf{HE}\wedge\neg\mathsf{A}\wedge\mathsf{E} & \neg\mathsf{HE}\wedge\mathsf{A}\wedge\neg\mathsf{E} & \neg\mathsf{HE}\wedge\mathsf{A}\wedge\mathsf{E} \\ 0.14 & 0.006 & 0.156 & 0.018 \\ & & & \\ \mathsf{HE}\wedge\neg\mathsf{A}\wedge\neg\mathsf{E} & \mathsf{HE}\wedge\neg\mathsf{A}\wedge\mathsf{E} & \mathsf{HE}\wedge\mathsf{A}\wedge\neg\mathsf{E} & \mathsf{HE}\wedge\mathsf{A}\wedge\mathsf{E} \\ 0.14 & 0.014 & 0.364 & 0.162. \end{array} \tag{3.5}$$

With this additional information, it will certainly not be expected that in the thought experiment either no inexpensive or no American films will show up. Rather, this information suggests that the relative frequency of expensive films will not much exceed the assumed minimal value 0.7, since expensive films are rather rare, and that the number of American films will not be far below the assumed maximal value 0.5, since American films are very common, particularly among expensive productions.

A tentative prediction for the frequencies of combinations of the properties A and E might thus be given by

$$\begin{array}{cccc} \neg\mathsf{A}\wedge\neg\mathsf{E} & \neg\mathsf{A}\wedge\mathsf{E} & \mathsf{A}\wedge\neg\mathsf{E} & \mathsf{A}\wedge\mathsf{E} \\ 0.195 & 0.4 & 0.05 & 0.4, \end{array} \tag{3.6}$$

which yields a prediction of

$$0.195 \cdot 0.5 + 0.4 \cdot 0.7 + 0.05 \cdot 0.7 + 0.4 \cdot 0.9 = 0.7725$$

for the relative frequency of HE, and with that a subjective probability 0.7725 for HEf. Since the relative frequencies (3.6) can hardly be predicted with full confidence, we will probably still only be willing to assign to HEf a degree of belief consisting of an interval containing the value 0.7725, but with somewhat tighter bounds than [0.64,0.8]. Rather arbitrarily, we may argue that [0.72,0.79] is a reasonable degree of belief for HEf.

What is the rationale for using statistical information in the way described by these examples for the prediction of the outcome of a thought experiment? Recall that the thought experiment consists of a series of realizations of some random mechanism that is our imperfect model of the chain of events leading to the observed event. Our understanding of the random mechanism producing the toss of the die in example 3.2.1 is characterized by the assumption that a fair die has been tossed in a non-manipulative manner.

In the film-example the screening of that particular film at that particular time is perceived to be a random draw from the set of all screenings of mystery films by arbitrary networks at arbitrary times.

Thus, in both examples the random mechanism used as an explanation of the chain of events producing the observed event is *equivalent* to the sampling rule (cf. section 2.1) to which the statistical data refers – equivalent in the sense that when we consider an arbitrary series of realizations of the random mechanism, i.e. one in which it is not supposed that each realization supplies us with a specific sort of evidence, then we would predict that the relative frequencies in this series agree with the statistical data. This assumed equivalence of the random mechanism associated with the individual observed event on the one hand, and the sampling rule associated with the statistical data on the other, is the key (default-) assumption made in default reasoning about probabilities. This observation is formally codified in a second postulate.

Postulate 2: *Default reasoning about probabilities rests on the assumption that the observed event* e *is a realization of a random mechanism equivalent to the sampling rule on which the statistical information is based.*

The assumption described in this postulate usually will be an idealization that we are ready to make, but which we know is unlikely to be strictly true. In example 3.2.2 for instance, our statistical information about mystery films may not have been obtained by considering actual broadcasts of mystery films, but by compiling this information from a film guide – resulting in data that does not reflect the different likelihood of different films to be shown on TV. For this data, the assumption in postulate 2 obviously is not literally true, but we will perhaps be still willing to make it as an idealization in order to use the statistical information for estimating the likelihood of HE for the current film.

Postulate 2 only describes a precondition that must be fulfilled in order to combine degrees of belief derived from evidence with statistical information. It gives no hint whatsoever by what operational rule this combination will actually be performed. In the die example the available data was of such a simple form that the way to combine it (by Jeffrey's rule) seems perfectly obvious. In the film example the best prediction for the frequency of happy ends is not as easily obtained. With the information (3.1) it has only been possible to argue for some upper and lower bounds, but there is no compelling argument for more specific values derivable from the data. With the additional statistical information (3.4) we can find arguments to improve our previous bounds; however, there appears to be no elementary argument by which these improved bounds can be derived in a principled manner.

A key observation that will be instrumental for a derivation of a specific analytical rule to perform the combination of prior degrees of belief with statistical background knowledge can be made by reconsidering the arguments used above in deriving bounds for the probabilities of $\rho_1(\mathrm{t})$ and $\mathsf{HE}\mathrm{f}$: in both cases, the predictions for the relative frequencies of these properties in the thought experiments as, respectively, 0.1 and [0.64,0.8] (or [0.72,0.79]) were obtained by only arguing from the prior beliefs derived from the evidence, and from the statistical data, but were completely independent of the evidence itself. When from a prior subjective probability of 0.3 for $\rho(\mathrm{t})$ and the statistical data available for tosses of fair dice, a degree of belief of 0.1 is derived for $\rho_1(\mathrm{t})$, this is done by simply considering a random sample (obtained by the sampling rule of the statistical data!) of tosses of a die, in which the relative frequency of the property ρ happens to be 0.3. For this imaginary sample it is no longer necessary to assume that each of its elements occurs in a setting analogous to the one of the original toss.

Similarly in the film example: assume that the scene we have seen does not provide any more relevant information with respect to the actual film f having any of the properties A, E or HE. Then, in order to predict the relative frequency of HE in the thought experiment associated with f, an arbitrary sample of mystery films with less than one half American and more than 70% expensive productions will be considered. If the original film happens to be black and white, and we have no statistical information referring to the property of being black and white, then we will not assume that every film in the random sample is black and white too, this property being recognized as irrelevant.

To obtain a more precise notion of what it means that the given evidence does not provide any more relevant information with respect to properties for which statistical information is available, let $\phi_1, \phi_2, \ldots$ be the (finite or infinite) set of properties to which our statistical information refers. Let Φ be a (finite or infinite) set of degrees of belief referring to the ϕ_i, i.e. Φ is a set of constraints on the subjective probability measure on $\phi_1, \phi_2, \ldots$. We say that Φ *exhausts the evidence* if, based on the evidence alone, and without any statistical information, we are unable to assign degrees of belief to the ϕ_i any more specific than the ones in Φ. The way in which statistical data is used to define degrees of belief is now described in a third postulate.

Postulate 3: *If the default assumption of postulate 2 has been made, and Φ is a set of degrees of belief exhausting the evidence obtained about an event* e *with regard to the properties $\phi_1, \phi_2, \ldots$ for which statistical information is available, then the predicted frequency of a property ϕ_i in the thought experiment associated with* e *is calculated as the expected relative frequency of the property ϕ_i in a large random sample of events (obtained by the sampling rule underlying the statistics), given that the relative frequencies of the properties $\phi_1, \phi_2, \ldots$ in that sample is within the bounds prescribed by Φ.*

Again, this postulate, in order to make it more intuitively intelligible, has been formulated somewhat sloppily by only speaking of a large random sample, where, in fact, we are looking at the expected limiting frequency of the property ϕ_i as the sample size tends towards infinity. Moreover, the condition that the relative frequencies of properties $\phi_1, \phi_2, \ldots$ in any specific sample is within the bounds prescribed by Φ is slightly too strong: since the precise meaning of prior degrees of belief, too, only is a constraint on the limits these frequencies attain, for any fixed sample size the constraints in Φ are only assumed to be approximately satisfied.

Figure 3.3 is an illustration by means of the die-example of the result of the analysis here given of the meaning of degrees of belief and the process of their formation using statistical knowledge, an analysis summarized in our three postulates.

Postulate 3 still is descriptive rather than prescriptive: it remains open what relative frequencies we should expect in the random samples here described.

To solve this problem, in the next section the scenario described by postulate 3 is formalized in a statistical model. From this statistical model very strong results about the relative frequencies to be expected can be derived.

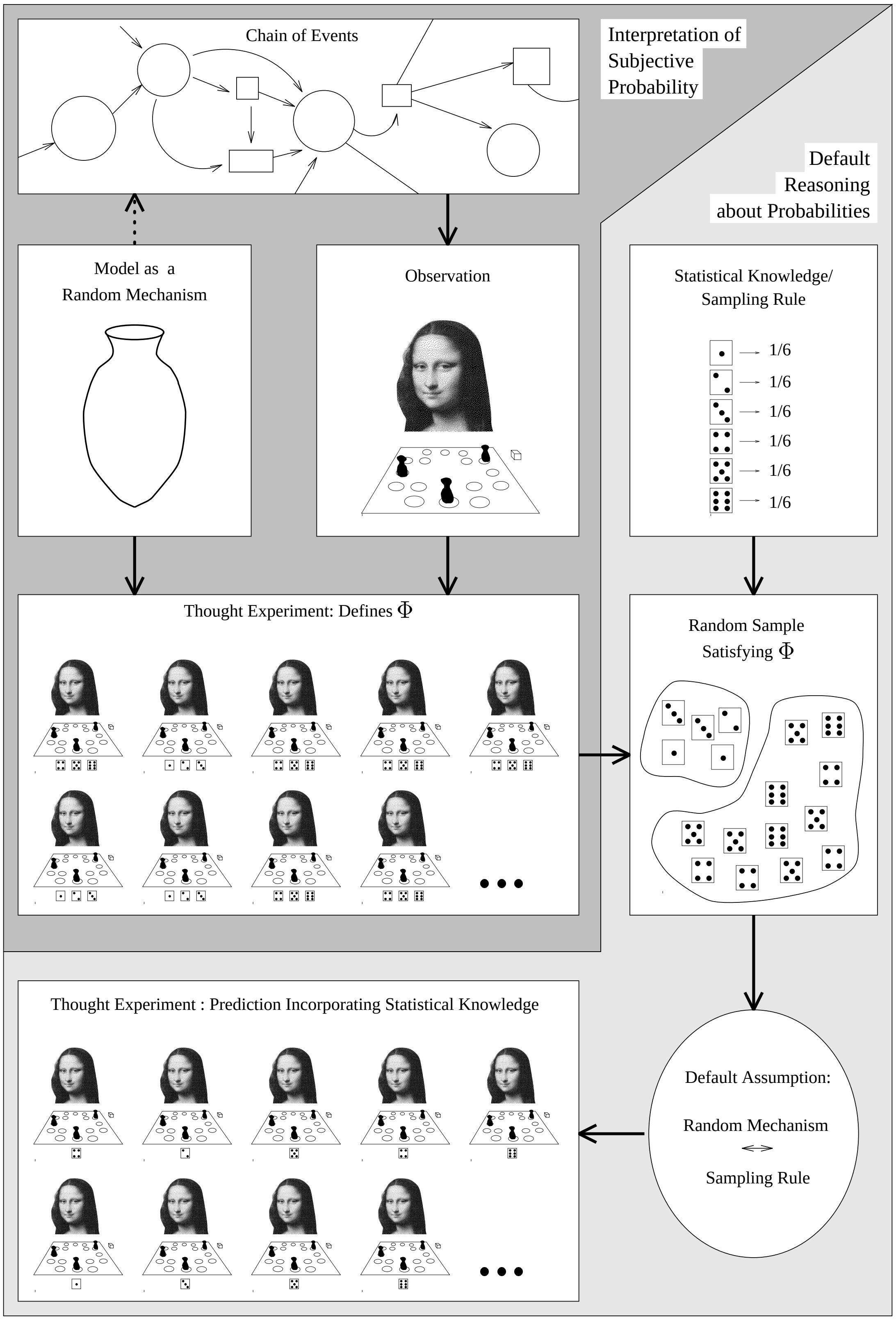

Figure 3.3: The analysis

3.4 The Statistical Model

3.4.1 Modeling Random Samples

In this section the results of the epistemic analysis contained in section 3.3 is implemented in a statistical model. The formalization will be carried out for the special case that events only are distinguished with regard to a finite set

$$\Gamma := \{\phi_1, \ldots, \phi_N\}$$

of properties, and that statistical probabilities and prior degrees of belief refer to these properties only. It will later be shown that results obtained for this case are sufficient for our subsequent project of developing a logical framework extending L^g_S incorporating formulation of, and reasoning with, subjective probabilities.

Without loss of generality, we may assume that the ϕ_i are mutually exclusive and exhaustive, i.e. each event possesses exactly one of the properties ϕ_i. If this is not the case, we can transform our statistical data to the 2^N mutually exclusive and exhaustive properties

$$\bigwedge_{i=1}^{N} \tilde{\phi}_i$$

with $\tilde{\phi}_i \in \{\phi, \neg\phi\}$.

Since within this framework events e and e$'$ are essentially indistinguishable iff $\phi_i(\mathrm{e})$ and $\phi_i(\mathrm{e}')$ for some $i \in \{1, \ldots, N\}$, we may actually regard Γ as the domain of possible events. The elements of Γ are the atoms of a boolean algebra with the operations of disjunction, conjunction, and negation (boolean algebras of this kind, of course, are known as Lindenbaum algebras). It does not present any difficulties to apply the definitions of a measure, which in section 1.2 has only been introduced for algebras of sets, to these more abstract boolean algebras. The set of measures on the Lindenbaum algebra generated by Γ, again, may be identified with Δ^N.

Statistical background information we possess then formulates constraints on a probability measure $\mu \in \Delta^N$, the probability measure belonging to the statistical sampling rule.

The subset of Δ^N containing the measures satisfying a set Φ of prior degrees of belief we denote by $\Delta(\Phi)$.

The mathematical model of a random sample of size n obtained by the statistical sampling rule now is given by a sequence $X_1, X_2, \ldots, X_n$ of independent random variables that are defined on some probability space Ω with probability σ-measure P, and take values in Γ according to the distribution μ, i.e.

$$\forall j, k \quad P^{X_j}(\phi_k) := P(\{\omega \in \Omega \mid X_j(\omega) = \phi_k\}) = \mu(\phi_k).$$

To model the limiting behaviour when larger and larger samples are drawn, we use an infinite sequence $X_1, X_2, \ldots$ of independent random variables with distribution $P^{X_j} = \mu$.

For each ϕ_k, the relative frequency of ϕ_k in the sample of size n is the random variable

$$\mathrm{P}^X_n(\phi_k) := \frac{1}{n} \sum_{j=1}^{n} 1_{\phi_k}(X_j),$$

with 1_{ϕ_k} the indicator variable of ϕ_k, i.e. $1_{\phi_k}(X_j(\omega)) = 1$ if $X_j(\omega) = \phi_k$, and 0 else. The tuple

$$\mathrm{P}_n^X := (\mathrm{P}_n^X(\phi_1), \dots, \mathrm{P}_n^X(\phi_N)) \in \Delta^N$$

then is a random variable on Ω: the *empirical distribution* of $X_1, \dots, X_n$.

What now has to be investigated, in order to calculate the degree of belief in ϕ_i as described in postulate 3, is the expected value of $\mathrm{P}_n^X(\phi_i)$ for large n, under the condition that P_n^X is within the bounds defined by the prior beliefs Φ. Recall, though, that postulate 3 has been formulated in a somewhat imprecise way in only referring to relative frequencies in large samples, rather than their limit as $n \to \infty$. Because of this, our last statement has to be refined: what must be examined is the limit for $n \to \infty$ of the expected value of P_n^X, given that P_n^X is close to $\Delta(\Phi)$.

3.4.2 Cross-Entropy

The fundamental tool for the characterization of the limiting behaviour of P_n^X turns out to be the *cross-entropy* of probability measures, which already has been mentioned in the introduction. Cross-entropy was first introduced by Kullback and Leibler ([1951]), and is also referred to as the Kullback-Leibler distance.

Cross-entropy (CE) is a function that maps pairs (ν, μ) of probability σ-measures into the extended set of positive reals:

$$CE: \quad \Delta_\sigma \mathfrak{A} \times \Delta_\sigma \mathfrak{A} \;\to \mathbf{R}^+ \cup \{\infty\}$$

with $\mathfrak{A}$ a σ-algebra over some set M and $\Delta_\sigma \mathfrak{A}$ the set of probability σ-measures on $\mathfrak{A}$.

For the definition of CE the probability-theoretic concept of absolute continuity is needed: for $\nu, \mu \in \Delta_\sigma \mathfrak{A}$ it is said that ν is *absolutely continuous* with respect to μ, written $\nu \ll \mu$, if

$$\forall A \in \mathfrak{A} \;\; \mu(A) = 0 \;\Rightarrow\; \nu(A) = 0.$$

By the Radon-Nikodym theorem, $\nu \ll \mu$ is equivalent to the existence of a density function f for ν with respect to μ (written $\nu = f\mu$), i.e. to the existence of a measurable function $f: \; M \to \mathbf{R}$ such that

$$\forall A \in \mathfrak{A} \;\; \nu(A) = \int_A f \mathrm{d}\mu.$$

Cross-entropy now can be defined by

$$CE(\nu, \mu) := \begin{cases} \infty & \text{if } \nu \not\ll \mu \\ \int f \ln f \, \mathrm{d}\mu & \text{if } \nu = f\mu. \end{cases} \tag{3.7}$$

When ν and μ are probability measures on a finite space, i.e. $\nu = (\nu_1, \dots, \nu_N), \mu = (\mu_1, \dots, \mu_N) \in \Delta^N$ for some N (and there is no difference between finite and σ-additivity), we may write

$$CE(\nu, \mu) := \begin{cases} \infty & \text{if } \exists i \in \{1, \dots, N\} \;\mu_i = 0 \wedge \nu_i > 0 \\ \displaystyle\sum_{\substack{i \in \{1,\dots,N\} \\ \mu_i > 0}} \nu_i \ln \frac{\nu_i}{\mu_i} & \text{else.} \end{cases} \quad . \tag{3.8}$$

By a standard convention, terms $0 \ln 0$ that may appear in the given sum are defined to equal 0.

Cross entropy often is referred to as a "measure of the distance between two probability measures" [Diaconis and Zabell, 1982], or a "measure of information dissimilarity for two probability measures" [Shore, 1986]. These interpretations have to be taken with a grain of salt, however. Note in particular that neither is CE symmetric nor does it satisfy the triangle inequality. All that CE has in common with a metric is positivity:

$$CE(\nu,\mu) \geq 0, \tag{3.9}$$

where equality holds iff $\nu = \mu$ ([Kullback and Leibler, 1951]).

A generalization of cross-entropy to a function on $2^{\Delta_\sigma \mathfrak{A}} \times \Delta_\sigma \mathfrak{A}$ is commonly found in the statistical literature: for $J \subseteq \Delta_\sigma \mathfrak{A}$, $\mu \in \Delta_\sigma \mathfrak{A}$ define

$$CE(J,\mu) := \mathit{inf}\,\{CE(\nu,\mu) \mid \nu \in J\}. \tag{3.10}$$

Of greatest interest usually are elements $\nu \in J$ that minimize $CE(\cdot,\mu)$ within J, i.e. that satisfy $CE(\nu,\mu) = CE(J,\mu)$. The interpretation of CE as a distance function gives rise to the conception of such ν as the projections of μ into J, which motivates the notation introduced by the following definition.

Definition 3.4.1 Let $\mathfrak{A}$ be a σ-algebra, $\mu \in \Delta_\sigma \mathfrak{A}$, $J \subseteq \Delta_\sigma \mathfrak{A}$. Define

$$\Pi_J(\mu) := \begin{cases} \emptyset & \text{if } CE(J,\mu) = \infty \\ \{\nu \in J \mid CE(\nu,\mu) = CE(J,\mu)\} & \text{else.} \end{cases} \tag{3.11}$$

In the special case that $\Pi_J(\mu)$ contains exactly one element, this element is denoted $\pi_J(\mu)$.

Many elementary properties of the mapping $(J,\mu) \mapsto \Pi_J(\mu)$ are derivable from the fact that for fixed μ, $CE(\cdot,\mu)$ is a strictly convex function on $\{\nu \in \Delta_\sigma \mathfrak{A} \mid CE(\nu,\mu) < \infty\}$, i.e. for any two different measures ν, ν' in this set and any $\lambda \in]0,1[$:

$$CE(\lambda\nu + (1-\lambda)\nu', \mu) < \lambda CE(\nu,\mu) + (1-\lambda) CE(\nu',\mu).$$

Some of these properties are collected in the following examples. Since in the sequel we will exclusively employ cross-entropy on finite spaces, the examples are formulated for this case only, even though analogous results hold in general.

Example 3.4.2 Let $J \subseteq \Delta^N$, $\mu \in \Delta^N$. If $\mu \in J$, then clearly $\Pi_J(\mu) = \{\mu\}$ because of the positivity property (3.9) of CE.

Suppose that $\mu \notin J$, and let $\nu \in \mathit{int}\, J$ with $CE(\nu,\mu) < \infty$. For some $\lambda > 0$ then $\lambda\mu + (1-\lambda)\nu \in J$. Since $CE(\mu,\mu) < CE(\nu,\mu)$, by the strict convexity of $CE(\cdot,\mu)$, $CE(\lambda\mu + (1-\lambda)\nu,\mu) < CE(\nu,\mu)$. Hence $\nu \notin \Pi_J(\mu)$, which shows that $\Pi_J(\mu) \subseteq \mathit{bd}\, J$.

Example 3.4.3 Let $J \subseteq \Delta^N$ be open, $\mu \notin J$. By the previous example it follows that $\Pi_J(\mu) = \emptyset$.

Example 3.4.4 Let $J \subseteq \Delta^N$ be closed, $\mu \in \Delta^N$ with $CE(J,\mu) < \infty$. Let $\nu_1, \nu_2, \ldots$ be a sequence in J with

$$CE(\nu_i, \mu) \to CE(J, \mu).$$

Since J, as a subset of the bounded set Δ^N, is bounded, this sequence may be assumed to converge to some $\nu \in \Delta^N$, which by the closedness of J belongs to J. By the continuity of $CE(\cdot, \mu)$, we have $lim\, CE(\nu_i, \mu) = CE(\nu, \mu) = CE(J, \mu)$, so that $\nu \in \Pi_J(\mu) \neq \emptyset$.

Example 3.4.5 Let $J \subseteq \Delta^N$ be closed and convex, $\mu \in \Delta^N$ with $CE(J,\mu) < \infty$. By the previous example, $\Pi_J(\mu)$ here is nonempty. By the convexity of $CE(\cdot, \mu)$, $\Pi_J(\mu)$ contains a unique element, because if $\nu, \nu' \in \Pi_J(\mu)$, $\nu \neq \nu'$, then $\nu'' := 1/2\nu + 1/2\nu' \in J$, and

$$CE(\nu'', \mu) < 1/2CE(\nu, \mu) + 1/2CE(\nu', \mu) = CE(\nu, \mu),$$

a contradiction.

Many interesting properties of cross-entropy minimization of a somewhat more sophisticated nature than exhibited in these four examples can be found in [Csiszár, 1975] and [Shore and Johnson, 1981].

3.4.3 The Limiting Behaviour of P_n^X

After these preparations, a very strong statement can be made about the limiting behaviour of P_n^X in the theorem below. For the statement of the theorem, some additional notation must be introduced.

For $\delta > 0$ and $\nu \in \Delta^N$ let

$$U_\delta(\nu) := \{\nu' \in \Delta^N \mid |\nu - \nu'| \leq \delta\}$$

denote the closed δ-ball around ν. Let

$$J(\delta) := \bigcup \{U_\delta(\nu) \mid \nu \in J\}$$

be the δ-hull around J.

For sequences $(\delta_n)_n, (\delta'_n)_n$ of real numbers we use the intuitive notation $(\delta_n) \geq 0$ to signify that $\delta_n \geq 0$ for all n, $(\delta_n) \geq (\delta'_n)$ for the fact that $\delta_n \geq \delta'_n$ for all n, and $(\delta_n) \searrow 0$ to say that $\delta_n \leq \delta_m$ for $n > m$, and $lim_{n\to\infty}\delta_n = 0$.

Theorem 3.4.6 Let $X_1, X_2, \ldots$ be a sequence of independent random variables taking values in $\{\phi_1, \ldots, \phi_N\}$ with distribution $\mu \in \Delta^N$. Let $J \subseteq \Delta^N$ be closed with $CE(J,\mu) < \infty$. Then there exists a sequence $(\delta_n) \searrow 0$, such that for all $(\delta'_n) \geq (\delta_n)$ with $(\delta'_n) \searrow 0$, there exists $(\epsilon_n) \searrow 0$, such that

$$lim_{n\to\infty} P(\mathrm{P}_n^X \in \Pi_J(\mu)(\epsilon_n) \mid \mathrm{P}_n^X \in J(\delta'_n)) = 1. \tag{3.12}$$

Before we turn to the proof of this theorem, which will then engage us for the rest of this section, we give some illustration of its meaning and discuss its relevance in the framework of our thought-experiment interpretation for degrees of belief.

Figure 3.4 illustrates the probability distribution of P_n^X in Δ^N for three different values of n. Darker shading in the picture represents greater probability for P_n^X to lie in the part of Δ^N thus marked. By the law of large numbers, the probability density for P_n^X gets more and more concentrated around its expected value μ.

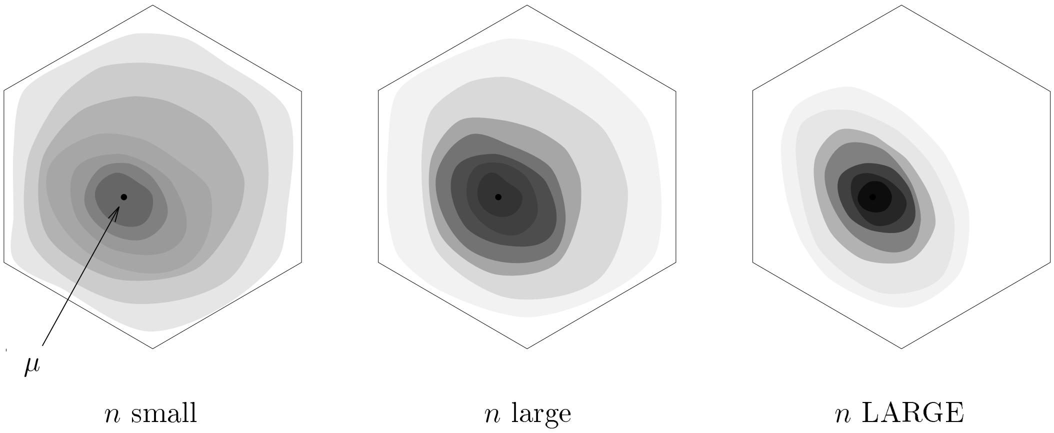

Figure 3.4: Distribution of P_n^X

Now consider Figure 3.5. Here attention has been restricted to values of P_n^X that lie in or near J. In figure 3.5 the special situation for a convex set J is depicted, so that $\Pi_J(\mu)$ here is a unique element $\pi_J(\mu)$, and $\Pi_J(\mu)(\epsilon_n) = U_{\epsilon_n}(\pi_J(\mu))$. With increasing n the distance δ'_n to J of points in Δ^N that are considered decreases. Theorem 3.4.6 now states that when we let δ'_n decrease sufficiently slowly, we can find for each n some $\epsilon_n \geq 0$ with $\epsilon_n \to 0$, so that the probability density of P_n^X in $J(\delta'_n)$ gets, in the limit, completely concentrated in $U_{\epsilon_n}(\pi_J(\mu))$.

A few comments may be useful to better understand the role that in theorem 3.4.6 is played by the sequence (δ_n). When J has interior points, then the theorem actually is true for $(\delta_n) = 0$, i.e. the whole process of approximating J by the sequence $J(\delta'_n)$ can be done without.

On the other hand, consider $J_r := \{(\nu_1, \ldots, \nu_N) \in \Delta^N \mid \nu_1 = r\}$ where r is some irrational number. Then P_n^X can never take a value in J, each component of P_n^X being of the form q/n for some $q \in \mathbf{N}$. Hence, conditioning on $\mathrm{P}_n^X \in J_r$ in (3.12) would mean to condition on the empty set, and the conditional probability in (3.12) would be undefined for every n. Moreover, for each n, P_n^X can only take on finitely many values, so that for sufficiently fast decreasing sequences $(\delta_n) > 0$, even $\{\mathrm{P}_n^X \in J_r(\delta_n)\}$ will be empty for all n. Thus, the condition of (δ'_n) "slowly" tending to 0 in theorem 3.4.6 makes sure that "sufficiently many" possible values of P_n^X are in $J(\delta'_n)$.

How does theorem 3.4.6 resolve the question about what relative frequencies should be expected in a random sample as described by postulate 3 and formalized in section 3.4.1?

First consider the case that the set $\Delta(\Phi)$ of belief measures consistent with the prior degrees of belief Φ is closed, and that $\Pi_{\Delta(\Phi)}(\mu) = \{\pi_{\Delta(\Phi)}(\mu)\}$ is a unique element of $\Delta(\Phi)$.

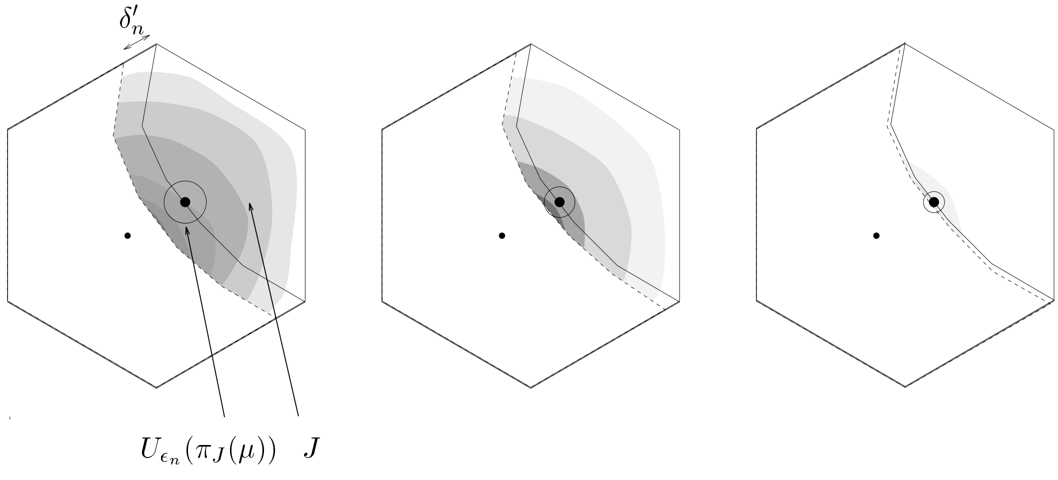

Figure 3.5: Distribution of P_n^X on $J(\delta_n)$

Theorem 3.4.6, in this situation, makes as strong a statement as one might wish for: when increasingly large random samples of events are considered that are assumed to satisfy the prior degrees of belief with (sufficiently slowly) increasing accuracy, then, in the limit, it is almost certain that the relative frequencies in the random sample drawn are defined by $\pi_{\Delta(\Phi)}(\mu)$.

In the case that $\Delta(\Phi)$ is closed, but $\Pi_{\Delta(\Phi)}(\mu)$ contains more than one element, theorem 3.4.6 still is immediately applicable to the situation described by postulate 3. The only difference now is that the relative frequencies in the random sample can only be predicted to be close to some member of $\Pi_{\Delta(\Phi)}(\mu)$. Theorem 3.4.6 does not provide any indication for some elements of $\Pi_{\Delta(\Phi)}(\mu)$ to define the relative frequencies in the random sample with greater probability than another.

The case of non-closed $\Delta(\Phi)$ gives rise to far less satisfactory results than the previous cases. This is not at all surprising, because even for very simple examples it is clear that default reasoning about probabilities runs into difficulties when the set of prior belief measures $\Delta(\Phi)$ is not closed: suppose, for instance, from the evidence observed a constraint $\nu(\phi) > 0.8$ has been derived for the belief measure ν of the property ϕ. Suppose, too, that the statistical probability of ϕ is $\mu(\phi) = 0.1$. Then the statistical information suggests to choose as small a value for $\nu(\phi)$ as permissible within the given constraint. Unfortunately, such a value does not exist. The probably most sensible way, in cases like this, to combine prior beliefs with statistical information would be to allow for a minor revision of prior beliefs, and, in the given example, to allow the value $\nu(\phi) = 0.8$ as the posterior subjective probability for ϕ.

How, then, do our epistemic model and its statistical formalization fare with non-closed prior belief sets Φ? By a closer look at postulate 3 – in its precise version with the imprecise statements about large samples substituted by the proper statements about the limiting behaviour as the sample size tends towards infinity – it is found that the method there described for combining prior degrees of belief with statistical knowledge does not really permit to distinguish between prior beliefs Φ and prior beliefs Ψ with $\Delta(\Psi) = cl\,\Delta(\Phi)$ at all: for both prior belief sets, for any fixed sample size, the same samples will be recognized as being in accordance with the prior beliefs, because any relative frequencies observed are close to $\Delta(\Phi)$ iff they are close to $cl\,\Delta(\Phi)$.

Thus, the rules described in postulate 3 according to which random samples are drawn and evaluated turn out to be the same for Φ and Ψ. Necessarily, then, the conclusions will also have to be the same that are derived for prior beliefs Φ and Ψ.

It is easy to show that this is actually what happens when (3.12) is generalized for non-closed $\Delta(\Phi)$: first, by letting $J := cl\,\Delta(\Phi)$,

$$lim_{n\to\infty} P(\mathrm{P}_n^X \in \Pi_{cl\,\Delta(\Phi)}(\mu)(\epsilon_n) \mid \mathrm{P}_n^X \in cl\,\Delta(\Phi)(\delta_n')) = 1 \tag{3.13}$$

is obtained. Since for all n with $\delta_n' > 0$

$$\Pi_{cl\,\Delta(\Phi)}(\mu) \subseteq cl\,\Delta(\Phi) \subseteq \Delta(\Phi)(\delta_n'),$$

in (3.13) the cl-operator can be dropped in the conditioning sets:

$$lim_{n\to\infty} P(\mathrm{P}_n^X \in \Pi_{cl\,\Delta(\Phi)}(\mu)(\epsilon_n) \mid \mathrm{P}_n^X \in \Delta(\Phi)(\delta_n')) = 1 \tag{3.14}$$

(for sufficiently slowly decreasing $(\delta_n') > 0$).

Thus, neither the epistemic, not the statistical model effectively distinguish between prior beliefs given by $\Delta(\Phi)$, and those given by $cl\,\Delta(\Phi)$.

One might wonder whether statements of the form (3.12) really are statistical results of the best possible form one might hope to obtain in order to compute expected frequencies according to postulate 3. A statement like

$$\mathrm{P}(lim_{n\to\infty}\mathrm{P}_n^X \in \Pi \mid lim_{n\to\infty}\mathrm{P}_n^X \in \Delta(\Phi)) = 1 \tag{3.15}$$

for some $\Pi \subseteq \Delta(\Phi)$, might be viewed as a more immediate formal statement about expected frequencies. Unfortunately, though, (3.15) does not make any sense because (unless $\mu \in \Delta(\Phi)$) the conditioning event $\{lim_{n\to\infty}\mathrm{P}_n^X \in \Delta(\Phi)\}$ has probability 0.

The conclusion we shall draw from this discussion is that for closed $\Delta(\Phi)$ theorem 3.4.6 permits the prediction that the relative frequencies in the random samples described by postulate 3 will approach a limit in $\Pi_{\Delta(\Phi)}(\mu)$ which therefore will define the result of combining prior beliefs Φ with statistical information μ. When $\Delta(\Phi)$ is non-closed there are essentially two possibilities: either we are willing to slightly revise prior beliefs, and adopt $\Pi_{cl\,\Delta(\Phi)}(\mu)$ as the set of posterior belief measures, or we insist on the satisfaction of the prior beliefs, in which case only $\Pi_{cl\,\Delta(\Phi)}(\mu) \cap \Delta(\Phi)$ will be admissible posterior belief measures. Should this last set be empty, then Φ can not consistently be combined with the statistical information μ.

We end this discussion of theorem 3.4.6 by looking at the results we obtain when applying this result to the examples of section 3.3. In the discussion of these examples we shall make use of the well-known fact that cross-entropy minimization is equivalent to Jeffrey's rule in those special cases where this rule is applicable (cf. [Diaconis and Zabell, 1982] for instance, and corollary 4.0.20 below).

In example 3.2.1, we have $\Gamma = \{\rho_1, \ldots, \rho_6\}$,

$$\mu = (1/6, \ldots, 1/6), \quad \Delta\Phi = \{(\nu_1, \ldots, \nu_6) \in \Delta^6 \mid \nu_1 + \nu_2 + \nu_3 = 0.3\}.$$

By application of Jeffrey's rule:

$$\pi_{\Delta\Phi}(\mu) = (0.1, 0.1, 0.1, 0.7/3, 0.7/3, 0.7/3).$$

Hence, the predicted frequencies of properties ρ_i, and thus the degrees of belief assigned to ρ_it are 0.1 for $i = 1, 2, 3$ and 0.7/3 for $i = 4, 5, 6$ in accordance with the intuitive reasoning in example 3.2.1.

For example 3.2.2 let

$$\Gamma := \{\neg\mathsf{HE} \wedge \neg\mathsf{A} \wedge \neg\mathsf{E}, \ldots, \mathsf{HE} \wedge \mathsf{A} \wedge \mathsf{E}\}.$$

Then

$$\Delta(\Phi) = \{\nu \in \Delta\Gamma \mid \nu(\mathsf{A}) \leq 0.5, \nu(\mathsf{E}) \geq 0.7\}$$

is closed and convex. The statistical distribution on Γ only is partially determined by the constraints (3.1). In fact, every distribution on $\neg\mathsf{A} \wedge \neg\mathsf{E}, \ldots, \mathsf{A} \wedge \mathsf{E}$, extended to Γ by the given conditional probabilities of HE, is consistent with (3.1).

First consider a statistical distribution μ with

$$\mu(\mathsf{E}) = 1 \quad \text{and} \quad \mu(\mathsf{A}) = 0.5.$$

Since cross-entropy minimizing measures $\nu \in \Delta(\Phi)$ must satisfy $\nu \ll \mu$, this implies that for such a ν certainly $\nu(\neg\mathsf{E}) = 0$ must also hold, so the whole minimization problem is reduced to the four element probability space $\Gamma' \subset \Gamma$ containing the elements with an unnegated conjunct E. $\Delta(\Phi)$ then reduces to the condition $\nu(\mathsf{A}) \leq 0.5$. Then we have $\mu \in \Delta(\Phi)$, and thus $\pi_{\Delta(\Phi)}(\mu) = \mu$, so that, as computed in (3.2),

$$\pi_{\Delta(\Phi)}(\mu)(\mathsf{HE}) = 0.8.$$

Now let μ' be defined by (3.1) and

$$\mu'(\mathsf{A}) = 0 \quad \text{and} \quad \mu'(\mathsf{E}) = 0.7.$$

With this μ' the minimization problem again is reduced to a four-element probability space, and the single remaining constraint $\nu(\mathsf{E}) \geq 0.7$ is satisfied by μ'. Hence, $\pi_{\Delta(\Phi)}(\mu') = \mu'$, and as in (3.3)

$$\pi_{\Delta(\Phi)}(\mu')(\mathsf{HE}) = 0.64.$$

Thus, the lower and upper bound derived by intuitive reasoning in section 3.3 correspond to the precise result theorem 3.4.6 applied under two different assumptions for the underlying statistical measure. For any other statistical measure μ'' on Γ that is consistent with (3.1), the value of $\pi_{\Delta(\Phi)}(\mu'')(\mathsf{HE})$ also will belong to the interval [0.64,0.8]. Hence, the statistical model yields the same result as the informal arguments in section 3.3.

When (3.4) is added to our statistical data, and a unique statistical distribution $\bar{\mu}$ not belonging to $\Delta(\Phi)$ is given by (3.5), the projection $\pi_{\Delta(\Phi)}(\bar{\mu})$ no longer is computable by elementary arguments from positivity. Instead, as in all general cross-entropy minimization problems, a non-linear optimization algorithm will have to be used, which only yields an approximation of $\pi_{\Delta(\Phi)}(\bar{\mu})$ (see [Wen, 1988] for one example of such an algorithm). Here we obtain

$$\pi_{\Delta(\Phi)}(\bar{\mu})(\mathsf{HE}) \approx 0.739. \tag{3.16}$$

Unlike in the first two cases, here by utilizing (3.12) for the prediction of the relative frequency of HE, we are able to derive a very specific result, where our intuitive reasoning only led to bounds [0.72,0.79] with an uncertain justification.

We now turn to the proof of theorem 3.4.6, which we break up into a series of separate theorems and lemmas.

The proof essentially rests on a result in the theory of large deviations from the sample mean. An important type of theorem obtained in this theory, the first instance of which was proved by Sanov [1957], provides a precise quantification of the exponential decay of the probability for the empirical distribution of a sequence of random variables to lie in certain subsets of probability measures: let $X_1, X_2, \ldots$ be a sequence of independent random variables taking values in $\{\phi_1, \ldots, \phi_N\}$ with distribution $\mu \in \Delta^N$. Let $J \subseteq \Delta^N$. We say that *the Sanov-theorem holds for* J, if $CE(J, \mu) < \infty$ and

$$lim_{n\to\infty} \frac{1}{n} \ln \mathrm{P}(\mathrm{P}_n^X \in J) = -CE(J, \mu). \tag{3.17}$$

Of course, (3.17) really makes a statement about $X_1, X_2, \ldots,\ \mu$, and J, rather than J alone as implied by the diction here introduced. In fact, different assumptions made about the space the X_i take their values in, their distribution, and the subsets J of probability measures on that space account for the great diversity of Sanov-theorems encountered in the literature. Being here solely concerned with random variables X_i taking values in $\{\phi_1, \ldots, \phi_N\}$ according to a specified measure μ, however, for us (3.17) becomes essentially a statement about J.

In the sequel, it is always assumed that $X_1, X_2, \ldots$ are as in theorem 3.4.6.

Statements of the form (3.17) are not yet quite what we need to prove theorem 3.4.6, because in that theorem limits of probabilities are considered that are defined by subsets of Δ^N dependent on n. We will therefore have to consider generalizations of (3.17): let $(J_\delta)_{\delta \in [0,\infty[}$ be a monotone family of subsets of Δ^N, i.e. $J_\delta \subseteq J_{\delta'}$ for $\delta \leq \delta'$ (the prime example of such a monotone family naturally being the set of δ-hulls $J(\delta)$ of some $J = J_0$). Let $(\delta_n) \searrow 0$. We say that *the Sanov-theorem holds for* $(J_{\delta_n})_n$ if $CE(J_0, \mu) < \infty$, and

$$lim_{n\to\infty} \frac{1}{n} \ln \mathrm{P}(\mathrm{P}_n^X \in J_{\delta_n}) = -CE(J_0, \mu). \tag{3.18}$$

For a monotone family $(J_\delta)_\delta$ in Δ^N define

$$\begin{aligned} c: \ [0,\infty[\ &\to [0,\infty] \\ \delta \ &\mapsto -CE(J_\delta, \mu). \end{aligned}$$

Lemma 3.4.7 Let $(J_\delta)_\delta$ be a monotone family of subsets of Δ^N. Assume that the Sanov-theorem holds for every J_δ with $\delta > 0$, and that c is continuous at $\delta = 0$. Then there exists $(\delta_n) \geq 0$ such that the Sanov-theorem holds for all $(J_{\delta'_n})_n$ with $(\delta'_n) \geq (\delta_n)$, $(\delta'_n) \searrow 0$.

Proof: Let $(\delta_i^0) \searrow 0$. Since the Sanov-theorem holds for every $J_{\delta_i^0}$, there exists for every i an index $k(i)$ such that for all $m > k(i)$:

$$\frac{1}{m}\ln\mathrm{P}(\mathrm{P}_m^X \in J_{\delta_i^0}) \in U_{1/i}(c(\delta_i^0)).$$

For each n let

$$\delta_n := min\{\delta_i^0 \mid n \geq max\{i, k(i)\}\} \geq \delta_n^0.$$

For some initial n the set $\{\delta_i^0 \mid n \geq max\{i, k(i)\}\}$ may be empty. For that case, define $min\,\emptyset := \infty$. The condition that $n \geq i$, in the definition of δ_n makes sure that $(\delta_n) \geq (\delta_n^0)$, which we shall use below.

By the definition, $(\delta_n) \searrow 0$, and because $c(\delta_n) \to c(0)$, the Sanov-theorem holds for (J_{δ_n}).

Now let $(\delta'_n) \geq (\delta_n)$ with $(\delta'_n) \searrow 0$. Then $\mathrm{P}(\mathrm{P}_n^X \in J_{\delta'_n}) \geq \mathrm{P}(\mathrm{P}_n^X \in J_{\delta_n})$, and thus

$$lim\ inf\ \frac{1}{n}\ln\mathrm{P}(\mathrm{P}_n^X \in J_{\delta'_n}) \geq c(0). \tag{3.19}$$

As for (δ_i^0) above, we can construct from (δ'_n) a sequence $(\delta''_n) \geq (\delta'_n)$ such that the Sanov-theorem holds for $(J_{\delta''_n})$. Thus

$$lim\ sup\ \frac{1}{n}\ln\mathrm{P}(\mathrm{P}_n^X \in J_{\delta'_n}) \leq lim\ sup\ \frac{1}{n}\ln\mathrm{P}(\mathrm{P}_n^X \in J_{\delta''_n}) = c(0). \tag{3.20}$$

(3.19) and (3.20) together mean that the Sanov theorem holds for $(J_{\delta'_n})$, thus proving the lemma. □

Next, a condition for ensuring continuity of c is formulated.

Lemma 3.4.8 Let $J_\delta \subseteq J(\delta)$ for all δ. Then c is continuous at $\delta = 0$.

Proof: Let $(\delta_n) \searrow 0$. Since $c(\delta) \leq c(0)$ for all $\delta \geq 0$, it is clear that

$$lim\ sup\ \ c(\delta_n) \leq c(0).$$

For each n let $\nu_n \in cl\ J_{\delta_n}$ with $CE(\nu_n, \mu) = c(\delta_n)$. Such ν_n always exist due to the boundedness of J_{δ_n}, and the continuity of $CE(\cdot, \mu)$. Consider a convergent subsequence $(\nu_{n_i})_i \subseteq (\nu_n)_n$. From the condition that $J_{\delta_n} \subseteq J(\delta_n)$ it follows that the limit ν of (ν_{n_i}) lies in $cl\ J$. Then (again utilizing the continuity of $CE(\cdot,\cdot)$)

$$\begin{aligned} lim\ inf_n\ c(\delta_n) &= lim\ inf_i\ c(\delta_{n_i}) \\ &= lim\ inf_i\ CE(\nu_{n_i}, \mu) \\ &= CE(\nu, \mu) \\ &\geq c(0). \end{aligned}$$

□

After the very general statements made in the preceding two lemmas, we can now approach the concrete situation dealt with in theorem 3.4.6. The following theorem is the basis for its proof.

Theorem 3.4.9 Let $J \subseteq \Delta^N$ with
$$J \subseteq cl(int\ J). \tag{3.21}$$
Then the Sanov-theorem holds for J.

Proof: See [Bahadur, 1971], example 5.4 and lemma 5.2. $\square$

The applicability of theorem 3.4.9 to our situation is established by the following lemma.

Lemma 3.4.10 **(a):** Let $J \subseteq \Delta^N$, $\delta > 0$, then (3.21) holds for $J(\delta)$.
(b): Let (3.21) hold for $J \subseteq \Delta^N$. Let $A \subseteq \Delta^N$ be open. Then (3.21) holds for $J' := J \cap A$.

Proof: **(a):** For $J(\delta)$ it is even trivially true that $J(\delta) \subseteq int\ J(\delta)$. **(b):** The assertion is trivially true when $J' = \emptyset$. Let $\zeta \in J'$. For $\epsilon > 0$ let $U^0_\epsilon(\zeta)$ be the interior of $U_\epsilon(\zeta)$. It must be shown that for $\epsilon > 0$: $U^0_\epsilon(\zeta) \cap int\ J' \neq \emptyset$. From the openness of A it follows that $int\ J' = int\ J \cap A$, so that it must be shown that

$$U^0_\epsilon(\zeta) \cap int\ J \cap A \neq \emptyset$$

Since $\zeta \in J$, and (3.21) holds for J, we have that $U^0_\epsilon(\zeta) \cap int\ J \neq \emptyset$.

Now assume that $U^0_\epsilon(\zeta) \cap int\ J \subseteq A^c$. Using $E \cap cl F \subseteq cl(E \cap F)$ for open sets E, we get $U^0_\epsilon(\zeta) \cap cl(int\ J) \subseteq A^c$, and hence $U^0_\epsilon(\zeta) \cap J \subseteq A^c$. This last inclusion, however, contradicts $U^0_\epsilon(\zeta) \cap J \cap A \neq \emptyset$, which must hold because $\zeta \in J'$. Therefore $\emptyset \neq U^0_\epsilon(\zeta) \cap int\ J \not\subseteq A^c$, and thus $U^0_\epsilon(\zeta) \cap\ int\ J \cap A \neq \emptyset$. $\square$

This concludes the preparations for the proof of theorem 3.4.6, which now it is not hard to obtain.

Proof of theorem 3.4.6: Fix $\epsilon > 0$, and let $\delta > 0$. Abbreviate $\Pi_J(\mu)(\epsilon)$ by Π. Then the sets $J(\delta)$ and $J(\delta) \setminus \Pi$ are as in lemma 3.4.10 (a) and (b) respectively. Thus, by theorem 3.4.9, the Sanov-theorem holds for $J(\delta)$ and $J(\delta) \setminus \Pi$.

Both the two families $(J(\delta))_\delta$ and $(J(\delta)\setminus\Pi)_\delta$ satisfy the condition of lemma 3.4.8, so that by lemma 3.4.7 there exists $(\delta^*_n) \searrow 0$ and $(\delta^{**}_n) \searrow 0$ such that for every $(\overline{\delta}_n) \searrow 0$ with $(\overline{\delta}_n) \geq (\delta^*_n)$ the Sanov-theorem holds for $(J(\overline{\delta}_n))$, and for every $(\overline{\overline{\delta}}_n) \searrow 0$ with $(\overline{\overline{\delta}}_n) \geq (\delta^{**}_n)$ the Sanov-theorem holds for $(J(\overline{\overline{\delta}}_n) \setminus \Pi)$. Define $\delta_n := max\{\delta^*_n, \delta^{**}_n\}$. Clearly $(\delta_n) \searrow 0$, and for every $(\delta'_n) \geq (\delta_n)$ with $(\delta'_n) \searrow 0$ the Sanov-theorem holds for $(J(\delta'_n))$, and $(J(\delta'_n) \setminus \Pi)$. For such (δ'_n) then

$$\begin{aligned} lim_{n\to\infty}\frac{1}{n}\ln\frac{P(\mathrm{P}^X_n \in J(\delta'_n) \setminus \Pi)}{P(\mathrm{P}^X_n \in J(\delta'_n))} &= lim_{n\to\infty}\frac{1}{n}[\ln P(\mathrm{P}^X_n \in J(\delta'_n) \setminus \Pi) - \ln P(\mathrm{P}^X_n \in J(\delta'_n))] \\ &= CE(J,\mu) - CE(J \setminus \Pi, \mu). \end{aligned}$$

Consider $\nu \in cl\ (J \setminus \Pi)$ with $CE(\nu, \mu) = CE(J \setminus \Pi, \mu)$. Since J was assumed to be closed, we have $\nu \in J$. All points in $J \setminus \Pi$ having at least a Euclidean distance ϵ to points in $\Pi_J(\mu)$, furthermore $\nu \notin \Pi_J(\mu)$. Therefore $CE(J \setminus \Pi, \mu) > CE(J, \mu)$, so that

$$lim_{n\to\infty}\frac{1}{n}\ln\frac{P(\mathrm{P}^X_n \in J(\delta'_n) \setminus \Pi)}{P(\mathrm{P}^X_n \in J(\delta'_n))} < 0.$$

It follows that

$$lim_{n\to\infty}\ln\frac{P(\mathrm{P}_n^X \in J(\delta'_n)\setminus\Pi)}{P(\mathrm{P}_n^X \in J(\delta'_n))} = -\infty,$$

and hence

$$lim_{n\to\infty}\frac{P(\mathrm{P}_n^X \in J(\delta'_n)\setminus\Pi)}{P(\mathrm{P}_n^X \in J(\delta'_n))} = 0.$$

Thus

$$\begin{aligned} lim_{n\to\infty}P(\mathrm{P}_n^X \in \Pi \mid \mathrm{P}_n^X \in J(\delta'_n)) &= lim_{n\to\infty}\frac{P(\mathrm{P}_n^X \in J(\delta'_n)\cap\Pi)}{P(\mathrm{P}_n^X \in J(\delta'_n))} \\ &= lim_{n\to\infty}\left(1-\frac{P(\mathrm{P}_n^X \in J(\delta'_n)\setminus\Pi)}{P(\mathrm{P}_n^X \in J(\delta'_n))}\right) \\ &= 1. \end{aligned}$$

Since ϵ was arbitrary, we can find a sequence $(\epsilon_n) \searrow 0$ satisfying (3.12). □

Results related to theorem 3.4.6 are contained in [van Campenhout and Cover, 1981]. Here, too, results from large deviation theory are employed to characterize the limiting behaviour of certain conditional probabilities by means of cross-entropy – the conditioning sets, however, being defined somewhat differently from constraining the empirical distribution as we did here. In [Groeneboom *et al.*, 1979] several results can be found that are related to our lemma 3.4.7.

Also it should be noted that Sanov-theorems are a generalization of the well-known combinatorial concentration phenomenon at the maximum entropy distribution ([Jaynes, 1983],[Paris and Vencovská, 1989],[Grove *et al.*, 1992b]).

Chapter 4

Cross-Entropy in Real Closed Fields

The analysis of default reasoning about probabilities conducted in the previous chapter identified cross-entropy minimization as its central tool. This tool, so far, only is available for real-valued probabilities. In order to incorporate cross-entropy minimization into a logical framework based on probabilities in rc-fields, a cross-entropy function must be explained for general **F**-valued probability measures. This amounts to defining a logarithmic function on a real closed field $\mathfrak{F}$, cross-entropy then being given by a combination of the logarithm, addition, and multiplication.

The study of extensions of (ordered or real closed) fields by an exponential function is an active research area in model theory, mainly aiming at resolving a conjecture by Tarski that the theory of ($\mathbf{R}$, exp) is decidable (e.g. [Wolter, 1986], [Ressayre, 1993]).

While, generally, it is more practical to work with exponential functions than logarithmic functions – the former being total on $\mathbf{R}$, and therefore not requiring any considerations for elements outside the domain of the function – for our particular purpose a direct axiomatization of the logarithmic function seems to be more convenient.

Definition 4.0.11 Let $\mathrm{S_{LOF}} := \{0, 1, +, \cdot, \leq, \mathsf{ln}\}$. LRCF is the set of axioms comprising the axioms RCF for real-closed fields and the following axioms.

$$\forall x, y > 0 \;\; \mathsf{ln}(x \cdot y) = \mathsf{ln}(x) + \mathsf{ln}(y) \tag{FUN}$$

$$\forall x > 0 \;\; x \neq 1 \rightarrow \mathsf{ln}(x) < x - 1 \tag{BD}$$

These two axioms implicitly define the standard logarithm in $\mathbf{R}$, i.e. $(\mathbf{R}, \ldots, \mathsf{ln}^{\mathrm{R}}) \models \mathrm{FUN} \wedge \mathrm{BD}$ iff ln^{R} restricted to arguments $x > 0$ is the standard natural logarithm ln. (This somewhat tortured distinction between the natural logarithm ln on $\mathbf{R}$, and the interpretation ln^{R} in an $\mathrm{S_{LOF}}$-extension of $\mathfrak{R}$ that models LRCF only is necessary because of ln not being a total function; it would be redundant if we worked with the exponential function.)

An $\mathrm{S_{LOF}}$- structure that is a model of LRCF we call a *logarithmic real closed field*, or *lrc-field* for short. As in $(\mathfrak{R}, \mathsf{ln}^{\mathrm{R}})$, in such an lrc-field the interpretation of $\mathsf{ln}(x)$ will also have to be defined for arguments ≤ 0. Since the axioms do not make any prescriptions for such arguments, these interpretations may be completely arbitrary.

Lemma 4.0.12 The following sentences are derivable from LRCF.

$$\mathsf{ln}(1) = 0 \tag{4.1}$$
$$\forall x > 0 \;\; \mathsf{ln}(1/x) = -\mathsf{ln}(x) \tag{4.2}$$
$$\forall x \in (0,1) \;\; \mathsf{ln}(x) < 0 \tag{4.3}$$
$$\forall x > 1 \;\; \mathsf{ln}(x) > 0 \tag{4.4}$$
$$\forall x, y > 0 \;\; x < y \rightarrow \mathsf{ln}(x) < \mathsf{ln}(y) \tag{4.5}$$
$$0 \cdot \mathsf{ln}(0) = 0 \tag{4.6}$$

Proof: (4.1) is immediate from FUN; (4.2) follows from FUN and (4.1). (4.3) follows from BD; (4.4) is obtained from from (4.2) and (4.3).

To prove (4.5), write $y = \lambda \cdot x$ for some $\lambda > 1$. Then $\mathsf{ln}(y) = \mathsf{ln}(x) + \mathsf{ln}(\lambda) > \mathsf{ln}(x)$ by (4.4).

(4.6) is trivial, because $\forall x \; 0 \cdot x = 0$ is given by RCF. □

We now turn to the generalized cross-entropy function defined in an lrc-field $\mathfrak{F}$. With the logarithm function ln^{F} at our disposal, the definition of cross-entropy in $\mathbf{R}$ by (3.8) can basically be copied for defining cross-entropy in $\mathfrak{F}$.

In order to later be able to axiomatize the general cross-entropy definition in first-order logic, however, a small modification has to be made. We may not use a special ideal element ∞ assumed to be greater than every field element as the value for $CE(\nu, \mu)$ when $\nu \not\ll \mu$, because such an element is not contained in $\mathfrak{F}$ itself. For this reason, the definition of CE has to be restricted to pairs of measures $(\nu, \mu) \in \Delta^N_{\mathrm{F}} \times \Delta^N_{\mathrm{F}}$ with $\nu \ll \mu$. For such $\nu = (\nu_1, \ldots, \nu_N), \mu = (\mu_1, \ldots, \mu_N)$ let

$$CE^{\mathrm{F}}(\nu, \mu) := \sum_{\substack{i \in \{1,\ldots,N\} \\ \mu_i > 0}} \nu_i \mathsf{ln}^{\mathrm{F}} \frac{\nu_i}{\mu_i}. \tag{4.7}$$

Note that CE^{F} only is a function defined "from outside" for the semantical structure $\mathfrak{F}$. There is no new function symbol CE introduced, that then is interpreted inside $\mathfrak{F}$ by an extension CE^{F}. Proceeding by such a course would in fact require to introduce function symbols CE_N for all $N \geq 1$ with arities $2N$, only to be applied to probability measures of the appropriate size, and interpreted for all possible arguments, not only for ν, μ with $\nu \ll \mu$.

The following lemma states a key analytical property of the logarithmic function that is instrumental for the proof of important characteristics of cross-entropy and cross-entropy minimization.

Lemma 4.0.13 It is derivable from LRCF that

$$\forall x_1, y_1, x_2, y_2 > 0: \quad x_1 \mathsf{ln}\left(\frac{x_1}{y_1}\right) + x_2 \mathsf{ln}\left(\frac{x_2}{y_2}\right) \geq (x_1 + x_2) \mathsf{ln}\left(\frac{x_1 + x_2}{y_1 + y_2}\right), \tag{4.8}$$

where equality holds iff

$$\frac{x_1}{x_1 + x_2} = \frac{y_1}{y_1 + y_2}. \tag{4.9}$$

Proof: We make the substitutions

$$x := x_1 + x_2,$$
$$y := y_1 + y_2.$$

Then, for suitable $\lambda_x, \lambda_y \in]0,1[$:

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

and the left hand side of (4.8) may be rewritten as

$$\begin{aligned} & \lambda_x x \ln\left(\frac{\lambda_x x}{\lambda_y y}\right) + (1-\lambda_x)x\ln\left(\frac{(1-\lambda_x)x}{(1-\lambda_y)y}\right) \\ = \; & x\ln\left(\frac{x}{y}\right) + x\left(\lambda_x \ln\left(\frac{\lambda_x}{\lambda_y}\right) + (1-\lambda_x)\ln\left(\frac{1-\lambda_x}{1-\lambda_y}\right)\right). \end{aligned} \tag{4.10}$$

The equality condition (4.9) is equivalent to $\lambda_x = \lambda_y$, in which case the second term of (4.10) vanishes by (4.1), so that (4.8) holds with equality.

Now suppose that $\lambda_x \neq \lambda_y$. Then $\frac{\lambda_y}{\lambda_x} \neq 1$ and $\frac{1-\lambda_y}{1-\lambda_x} \neq 1$. Since $-\ln(x) > 1 - x$ for $x \neq 1$ by BD, we obtain

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

Since $x > 0$, this means that the second term of (4.10) is strictly greater than 0. This proves the lemma.

□

The first important application of lemma 4.0.13 is the following.

Lemma 4.0.14 (Positivity) Let $\mathfrak{F}$ be an lrc-field, $N \geq 2$, $\nu, \mu \in \Delta_{\mathrm{F}}^N$ with $\nu \ll \mu$. Then $CE^{\mathrm{F}}(\nu,\mu) \geq 0$ with equality iff $\nu = \mu$.

Proof: By induction on N. Let $N = 2$, $\nu = (\nu_1, \nu_2), \mu = (\mu_1, \mu_2) \in \Delta_{\mathrm{F}}^2$, $\nu \ll \mu$. If one of the μ_i equals 0, then so does the corresponding ν_i, in which case $\nu = \mu$ and $CE^{\mathrm{F}}(\nu,\mu) = 1\ln(1) = 0$. Suppose, then, that $\mu_i > 0$ $(i = 1,2)$. If $\nu_i = 0$ for one i, say $i = 1$, then $\nu \neq \mu$ and $CE^{\mathrm{F}}(\nu,\mu) = \ln\left(\frac{1}{\mu_2}\right) > 0$ by (4.4).

For the case that $\nu_i, \mu_i > 0$ $(i = 1,2)$, we have

$$\begin{aligned} CE^{\mathrm{F}}(\nu,\mu) &= \nu_1 \ln^{\mathrm{F}}(\frac{\nu_1}{\mu_1}) + \nu_2 \ln^{\mathrm{F}}(\frac{\nu_2}{\mu_2}) \\ &\geq (\nu_1+\nu_2)\ln^{\mathrm{F}}(\frac{\nu_1+\nu_2}{\mu_1+\mu_2}) \\ &= 1\ln^{\mathrm{F}}(1) \;=\; 0 \end{aligned}$$

by lemma 4.0.13 and (4.2), with equality iff $\nu_1/(\nu_1+\nu_2) = \mu_1/(\mu_1+\mu_2)$, i.e. $\nu = \mu$.

Now let $N > 2$, and assume that the lemma has been shown for $N-1$. For $\nu = \mu$ we again obtain $CE^{\mathrm{F}}(\nu, \mu) = 1\mathsf{ln}^{\mathrm{F}}(1) = 0$. Suppose, then, that $\nu \neq \mu$. Without loss of generality, $\nu_1 \neq \mu_1$. Define $\bar{\nu}, \bar{\mu} \in \Delta_{\mathrm{F}}^{N-1}$ by

$$\bar{\nu}_i := \nu_i \quad \bar{\mu}_i := \mu_i \qquad i = 1, \ldots, N-2,$$

and

$$\bar{\nu}_{N-1} := \nu_{N-1} + \nu_N \quad \bar{\mu}_{N-1} := \mu_{N-1} + \mu_N.$$

Then $\bar{\nu} \ll \bar{\mu}$, $\bar{\nu} \neq \bar{\mu}$, so that by induction hypothesis $CE^{\mathrm{F}}(\bar{\nu}, \bar{\mu}) > 0$. By lemma 4.0.13 we have $CE^{\mathrm{F}}(\nu, \mu) \geq CE^{\mathrm{F}}(\bar{\nu}, \bar{\mu})$, which proves the lemma. □

Lemma 4.0.15 (Convexity) Let $\mathfrak{F}$ be an lrc-field, $N \geq 2$, $\nu, \nu', \mu \in \Delta_{\mathrm{F}}^N$, $\nu \neq \nu'$ with $\nu, \nu' \ll \mu$. Let $0^{\mathrm{F}} < \lambda < 1^{\mathrm{F}}$. Then

$$CE^{\mathrm{F}}(\lambda\nu + (1-\lambda)\nu', \mu) < \lambda CE^{\mathrm{F}}(\nu, \mu) + (1-\lambda) CE^{\mathrm{F}}(\nu', \mu).$$

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

$$c_y: \quad x \mapsto x\mathsf{ln}^{\mathrm{F}}\left(\frac{x}{y}\right)$$

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

$$\begin{aligned} CE^{\mathrm{F}}(\lambda\nu + (1-\lambda)\nu', \mu) &= \sum_{\mu_i > 0} c_{\mu_i}(\lambda\nu_i + (1-\lambda)\nu'_i) \\ &< \sum_{\mu_i > 0} \lambda c_{\mu_i}(\nu_i) + (1-\lambda) c_{\mu_i}(\nu'_i) \\ &= \lambda CE^{\mathrm{F}}(\nu, \mu) + (1-\lambda) CE^{\mathrm{F}}(\nu', \mu), \end{aligned}$$

where the strict inequality holds because $\nu_i \neq \nu'_i$ for at least one $i \in \{1, \ldots, N\}$ with $\mu_i > 0$.

For the proof of the convexity of c_y, let $y > 0^{\mathrm{F}}$, $x_1, x_2 \geq 0^{\mathrm{F}}$, $x_1 \neq x_2$, $0^{\mathrm{F}} < \lambda < 1^{\mathrm{F}}$. 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^{F}, e.g. $x_1 = 0^{\mathrm{F}}$. Then

$$\begin{aligned} c_y(\bar{x}) &= (1-\lambda)x_2\mathsf{ln}^{\mathrm{F}}\left(\frac{(1-\lambda)x_2}{y}\right) \\ &< (1-\lambda)x_2\mathsf{ln}^{\mathrm{F}}\left(\frac{x_2}{y}\right) \\ &= \lambda c_y(x_1) + (1-\lambda)c_y(x_2), \end{aligned}$$

where the inequality is due to (4.5), and the final equality holds because $c_y(0) = 0$ by (4.6).

Now suppose that $x_1, x_2 > 0$. By lemma 4.0.13 we obtain

$$c_y(\bar{x}) \leq \lambda x_1\mathsf{ln}^{\mathrm{F}}\left(\frac{\lambda x_1}{y/2}\right) + (1-\lambda)x_2\mathsf{ln}^{\mathrm{F}}\left(\frac{(1-\lambda)x_2}{y/2}\right) \tag{4.11}$$

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

$$\lambda x_1 = (1-\lambda)x_2. \tag{4.12}$$

The right side of (4.11) may be rewritten as

$$\lambda x_1 \mathsf{ln}^{\mathrm{F}}\left(\frac{x_1}{y}\right) + \lambda x_1 \mathsf{ln}^{\mathrm{F}}(2\lambda) + (1-\lambda)x_2 \mathsf{ln}^{\mathrm{F}}\left(\frac{x_2}{y}\right) + (1-\lambda)x_2 \mathsf{ln}^{\mathrm{F}}(2(1-\lambda)).$$

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

$$c_y(\bar{x}) \leq \lambda c_y(x_1) + (1-\lambda)c_y(x_2) + \lambda x_1 \mathsf{ln}^{\mathrm{F}}(4\lambda(1-\lambda)), \tag{4.13}$$

still with equality iff (4.12) holds.

First consider the case that (4.12) in fact is true. Then, because $x_1 \neq x_2$, we have that $\lambda \neq 1/2$. By the completeness of the theory of RCF, and the fact that

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

we may infer that $4\lambda(1-\lambda) < 1$, which (with (4.3)) entails that $\lambda x_1 \mathsf{ln}^{\mathrm{F}}(4\lambda(1-\lambda)) < 0$, thus proving that

$$c_y(\bar{x}) < \lambda c_y(x_1) + (1-\lambda)c_y(x_2). \tag{4.14}$$

In almost the same manner (4.14) is derived for the case that (4.12) does not hold: the last term in (4.13) then is known to be ≤ 0, which suffices to prove (4.14) because we have strict inequality in (4.13). □

The next lemma states a very useful property of cross-entropy that plays the key role in the proof of the subset independence property of cross-entropy minimization shown by Shore and Johnson [1980] (cf. theorem 4.0.19 below). It has not been given a name by Shore and Johnson, or, apparently, elsewhere, so that we are free to call it the decomposition property.

Lemma 4.0.16 (Decomposition) Let $\mathfrak{F}$ be an lrc-field, $N \geq 2$, $\nu, \mu \in \Delta^N_{\mathrm{F}}$ with $\nu \ll \mu$. Let $P = \{P_1, \ldots, P_L\}$ be a partition of $\{1, \ldots, N\}$ with $P_h = \{i_{h,1}, \ldots, i_{h,n(h)}\} \subseteq \{1, \ldots, N\}$ $(h = 1, \ldots, L)$.

Let $\bar{\nu}, \bar{\mu} \in \Delta^L_{\mathrm{F}}$ be the restrictions of ν and μ to P, i.e.

$$\bar{\nu}_h = \sum_{g=1}^{n(h)} \nu_{i_{h,g}}, \qquad \bar{\mu}_h = \sum_{g=1}^{n(h)} \mu_{i_{h,g}}, \qquad (h = 1, \ldots, L).$$

Also, for $h = 1, \ldots, L$ with $\bar{\nu}_h > 0$ let $\nu^h, \mu^h \in \Delta^{n(h)}_{\mathrm{F}}$ be the conditional distributions of ν, μ on P_h, i.e.

$$\nu^h_g = \frac{\nu_{i_{h,g}}}{\bar{\nu}_h}, \qquad \mu^h_g = \frac{\mu_{i_{h,g}}}{\bar{\mu}_h}, \qquad (g = 1, \ldots, n(h)).$$

Then

$$CE^{\mathrm{F}}(\nu, \mu) = CE^{\mathrm{F}}(\bar{\nu}, \bar{\mu}) + \sum_{\substack{h=1 \\ \bar{\nu}_h > 0}}^{L} \bar{\nu}_h CE^{\mathrm{F}}(\nu^h, \mu^h). \tag{4.15}$$

Proof: First observe that $\nu \ll \mu$ implies $\bar{\nu} \ll \bar{\mu}$, so that $\bar{\nu}_h > 0$ only when $\bar{\mu}_h > 0$, hence μ^h is well-defined. Furthermore, $\nu^h \ll \mu^h$. (4.15) now is obtained by a straightforward computation.

$$
\begin{aligned}
CE^{\mathrm{F}}(\nu,\mu) &= \sum_{h=1}^{L} \sum_{\substack{g=1 \\ \mu_{i_{h,g}}>0}}^{n(h)} \nu_{i_{h,g}} \mathsf{ln}^{\mathrm{F}}\left(\frac{\nu_{i_{h,g}}}{\mu_{i_{h,g}}}\right) \\
&= \sum_{\substack{h=1 \\ \bar{\nu}_h > 0}}^{L} \left(\sum_{\substack{g=1 \\ \mu_{i_{h,g}}>0}}^{n(h)} \bar{\nu}_h \frac{\nu_{i_{h,g}}}{\bar{\nu}_h} \mathsf{ln}^{\mathrm{F}}\left(\frac{\nu_{i_{h,g}}/\bar{\nu}_h}{\mu_{i_{h,g}}/\bar{\mu}_h}\frac{\bar{\nu}_h}{\bar{\mu}_h}\right)\right) \\
&= \sum_{\substack{h=1 \\ \bar{\nu}_h > 0}}^{L} \left(\sum_{\substack{g=1 \\ \mu_{i_{h,g}}>0}}^{n(h)} \bar{\nu}_h \nu_g^h \left(\mathsf{ln}^{\mathrm{F}}\left(\frac{\nu_g^h}{\mu_g^h}\right) + \mathsf{ln}^{\mathrm{F}}\left(\frac{\bar{\nu}_h}{\bar{\mu}_h}\right)\right)\right) \\
&= \sum_{\substack{h=1 \\ \bar{\nu}_h > 0}}^{L} \bar{\nu}_h \mathsf{ln}^{\mathrm{F}}\left(\frac{\bar{\nu}_h}{\bar{\mu}_h}\right) + \sum_{\substack{h=1 \\ \bar{\nu}_h > 0}}^{L} \bar{\nu}_h \sum_{\substack{g=1 \\ \mu_{i_{h,g}}>0}}^{n(h)} \nu_g^h \mathsf{ln}^{\mathrm{F}}\left(\frac{\nu_g^h}{\mu_g^h}\right) \\
&= CE^{\mathrm{F}}(\bar{\nu},\bar{\mu}) + \sum_{\substack{h=1 \\ \bar{\nu}_h > 0}}^{L} \bar{\nu}_h CE^{\mathrm{F}}(\nu^h,\mu^h)
\end{aligned}
$$

□

In section 3.4.2 the definition of $CE(\nu,\mu)$ was followed by the definition of $CE(J,\mu)$ for $J \subseteq \Delta^N$. For arbitrary lrc-fields $\mathfrak{F}$ this definition can not be repeated because $inf\{CE^{\mathrm{F}}(\nu,\mu) \mid \nu \in J\}$ may not exist in $\mathbf{F}$. This means that for the projection $\Pi_J(\mu)$ we here have to use a slightly less economical definition than provided by (3.11).

Definition 4.0.17 Let $\mathfrak{A}$ be a finite algebra, $\mu \in \Delta_{\mathrm{F}}\mathfrak{A}$, $J \subseteq \Delta_{\mathrm{F}}\mathfrak{A}$. Define

$$\Pi_J(\mu) := \{\nu \in J \mid \nu \ll \mu,\ \forall \nu' \in J\ CE^{\mathrm{F}}(\nu',\mu) \geq CE^{\mathrm{F}}(\nu,\mu)\}. \tag{4.16}$$

In the special case that $\Pi_J(\mu)$ contains exactly one element, this element is denoted $\pi_J(\mu)$.

Definition 4.0.17 coincides with definition 3.4.1, when applied to real-valued measures.

Of the four examples 3.4.2 - 3.4.5 only the first two remain to be valid in the context of lrc-fields, because in these only the positivity and convexity property of CE have been used. For examples 3.4.4 and 3.4.5, on the other hand, a convergence argument has been employed that makes use of the completeness of the real numbers and therefore fails in arbitrary real closed fields.

The following two theorems contain two key analytical properties of cross-entropy minimization. They have been called *system independence* and *subset independence* in [Shore and Johnson, 1980] where proofs are given for real-valued probability σ-measures (discrete or given by densities). The second of these properties, in particular, is of great practical importance, it being the most important tool for effectively computing $\pi_J(\mu)$ in special cases, without needing to employ an iterative optimization algorithm.

Theorem 4.0.18 (System Independence) Let $\mathfrak{A}, \mathfrak{A}'$ be finite algebras with atoms $\{A_1, \dots, A_p\}$ and $\{A'_1, \dots, A'_q\}$ respectively. Let $\mathfrak{F}$ be an lrc-field, $\mu \in \Delta^p_{\mathrm{F}}, \mu' \in \Delta^q_{\mathrm{F}},\ J \subseteq \Delta^p_{\mathrm{F}}, J' \subseteq \Delta^q_{\mathrm{F}}$. Define

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

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.

$$J^\times = \{\nu^\times \in \Delta_{\mathrm{F}}\mathfrak{A}^\times \mid \ \nu^\times \restriction_1 \mathfrak{A} \in J \ \wedge \ \nu^\times \restriction_2 \mathfrak{A}' \in J'\}.$$

Then

$$\Pi_{J^\times}(\mu^\times) = \Pi_J(\mu) \otimes \Pi_{J'}(\mu') := \{\nu \otimes \nu' \mid \nu \in \Pi_J(\mu), \nu' \in \Pi_{J'}(\mu')\}. \tag{4.17}$$

Proof: The proof is very similar to the one in [Shore and Johnson, 1980]. The fact that we are here dealing with **F**-valued probabilities makes little difference.

To prove the theorem, by direct computations it is shown that for $\nu^\times \in \Delta\mathfrak{A}^\times$ with marginal distribution ν on $\mathfrak{A}$ and ν' on $\mathfrak{A}'$

$$CE^{\mathrm{F}}(\nu^\times, \mu^\times) \geq CE^{\mathrm{F}}(\nu \otimes \nu', \mu^\times) \tag{4.18}$$

with equality only for $\nu^\times = \nu \otimes \nu'$, and that

$$CE^{\mathrm{F}}(\nu \otimes \nu', \mu^\times) = CE^{\mathrm{F}}(\nu, \mu) + CE^{\mathrm{F}}(\nu', \mu'). \tag{4.19}$$

From (4.18), (4.19) and the trivial observation

$$J \otimes J' := \{\nu \otimes \nu' \mid \nu \in J, \nu' \in J'\} \subseteq J^\times \tag{4.20}$$

the theorem then follows because for $\nu^\times \in J^\times$ with marginals ν and ν'

$$\begin{array}{rcll} \nu^\times \in \Pi_{J^\times}(\mu^\times) & \Leftrightarrow & \nu^\times = \nu \otimes \nu' \wedge \nu \otimes \nu' \in \Pi_{J^\times}(\mu^\times) & \text{by (4.18) and (4.20)} \\ & \Leftrightarrow & \nu^\times = \nu \otimes \nu' \wedge \nu \in \Pi_J(\mu) \wedge \nu' \in \Pi_{J'}(\mu') & \text{by (4.19) and (4.20)} \\ & \Leftrightarrow & \nu^\times \in \Pi_J(\mu) \otimes \Pi_{J'}(\mu') & \end{array}$$

For the proof of (4.18) use the denotations

$$\nu^\times_{ij} := \nu^\times(A_i \times A'_j), \qquad \nu_i := \nu(A_i) = \sum_{j=1}^{q} \nu^\times_{ij}, \qquad \nu'_j := \nu'(A'_j) = \sum_{i=1}^{p} \nu^\times_{ij}.$$

Then

$$\begin{aligned}\sum_{i,j} \nu_{ij}^{\times} \mathsf{ln}^{\mathrm{F}}(\mu_i \mu'_j) &= \sum_i \nu_i \mathsf{ln}^{\mathrm{F}}(\mu_i) + \sum_j \nu'_j \mathsf{ln}^{\mathrm{F}}(\mu'_j) \\ &= \sum_{i,j} \nu_i \nu'_j \mathsf{ln}^{\mathrm{F}}(\mu_i) + \sum_{i,j} \nu_i \nu'_j \mathsf{ln}^{\mathrm{F}}(\mu'_j) \\ &= \sum_{i,j} \nu_i \nu'_j \mathsf{ln}^{\mathrm{F}}(\mu_i \mu'_j),\end{aligned}$$

and analogously

$$\sum_{i,j} \nu_{ij}^{\times} \mathsf{ln}^{\mathrm{F}}(\nu_i \nu'_j) = \sum_{i,j} \nu_i \nu'_j \mathsf{ln}^{\mathrm{F}}(\nu_i \nu'_j).$$

Therefore

$$\begin{aligned}CE^{\mathrm{F}}(\nu^{\times}, \mu^{\times}) &= \sum_{i,j} \nu_{ij}^{\times} \mathsf{ln}^{\mathrm{F}}\left(\frac{\nu_{ij}^{\times}}{\mu_i \mu'_j}\right) \\ &= \sum_{i,j} \nu_{ij}^{\times} \mathsf{ln}^{\mathrm{F}}(\nu_{ij}^{\times}) - \sum_{i,j} \nu_{ij}^{\times} \mathsf{ln}^{\mathrm{F}}(\mu_i \mu'_j) \\ &= \sum_{i,j} \nu_{ij}^{\times} \mathsf{ln}^{\mathrm{F}}(\nu_{ij}^{\times}) - \sum_{i,j} \nu_{ij}^{\times} \mathsf{ln}^{\mathrm{F}}(\mu_i \mu'_j) - \sum_{i,j} \nu_{ij}^{\times} \mathsf{ln}^{\mathrm{F}}(\nu_i \nu'_j) + \sum_{i,j} \nu_{ij}^{\times} \mathsf{ln}^{\mathrm{F}}(\nu_i \nu'_j) \\ &= (\sum_{i,j} \nu_{ij}^{\times} \mathsf{ln}^{\mathrm{F}}(\nu_{ij}^{\times}) - \sum_{i,j} \nu_{ij}^{\times} \mathsf{ln}^{\mathrm{F}}(\nu_i \nu'_j)) + (\sum_{i,j} \nu_i \nu'_j \mathsf{ln}^{\mathrm{F}}(\nu_i \nu'_j) - \sum_{i,j} \nu_i \nu'_j \mathsf{ln}^{\mathrm{F}}(\mu_i \mu'_j)) \\ &= CE^{\mathrm{F}}(\nu^{\times}, \nu \otimes \nu') + CE^{\mathrm{F}}(\nu \otimes \nu', \mu^{\times}).\end{aligned}$$

By lemma 4.0.14, $CE^{\mathrm{F}}(\nu^{\times}, \nu \otimes \nu') \geq 0$, with equality only for $\nu^{\times} = \nu \otimes \nu'$.

(4.19) is verified as follows:

$$\begin{aligned}CE^{\mathrm{F}}(\nu \otimes \nu', \mu^{\times}) &= \sum_{i,j} \nu_i \nu'_j \mathsf{ln}^{\mathrm{F}}\left(\frac{\nu_i \nu'_j}{\mu_i \mu'_j}\right) \\ &= \sum_{i,j} \nu_i \nu'_j \mathsf{ln}^{\mathrm{F}}\left(\frac{\nu_i}{\mu_i}\right) + \sum_{i,j} \nu_i \nu'_j \mathsf{ln}^{\mathrm{F}}\left(\frac{\nu'_j}{\mu'_j}\right) \\ &= CE^{\mathrm{F}}(\nu, \mu) + CE^{\mathrm{F}}(\nu', \mu').\end{aligned}$$

□

Theorem 4.0.19 (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 $\mu \in \Delta_{\mathrm{F}}\mathfrak{A}$.

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 $\nu \in \Delta_{\mathrm{F}}\mathfrak{A}$ let $\bar{\nu}$ denote the restricted distribution $\nu \upharpoonright \bar{\mathfrak{A}}$, and ν^h the conditional of ν on $\mathfrak{A}^h$ $(h = 1, \ldots, L;\ \nu(A_h) > 0)$.

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

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

with $\bar{J}$ a set of constraints on $\bar{\nu}$, and J_h a set of constraints on ν^h. Precisely:

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

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

$$\nu^h \in \Pi_{J_h^*}(\mu^h). \tag{4.21}$$

Proof: The proof, again, is very similar to the one contained in [Shore and Johnson, 1980], consisting basically of an invocation of lemma 4.0.16.

Let $\nu \in \Pi_J(\mu)$ and $h \in \{1, \ldots, L\}$ with $\nu(A_h) > 0$. From $\nu \ll \mu$ it follows that $\mu(A_h) > 0$, so that both the conditional probability distributions in (4.21) exist.

Now suppose that there exists $\nu' \in J_h^*$ with

$$CE^{\mathrm{F}}(\nu', \mu^h) < CE^{\mathrm{F}}(\nu^h, \mu^h). \tag{4.22}$$

Then $\tilde{\nu} \in \Delta_{\mathrm{F}}\mathfrak{A}$ defined by

$$\begin{array}{ll} \bar{\tilde{\nu}} := \bar{\nu} & \\ \tilde{\nu}^k := \nu^k & \text{for } k \in \{1, \ldots, L\} \setminus \{h\},\ \nu(A_h) > 0 \\ \tilde{\nu}^h := \nu' & \end{array}$$

satisfies the constraints in J. By (4.15) and (4.22)

$$CE^{\mathrm{F}}(\nu, \mu) - CE^{\mathrm{F}}(\tilde{\nu}, \mu) = \nu(A_h)(CE^{\mathrm{F}}(\nu^h, \mu^h) - CE^{\mathrm{F}}(\nu', \mu^h)) > 0,$$

a contradiction. □

Corollary 4.0.20 (Jeffrey's Rule) Let $\mathfrak{A}$ be a finite algebra on M, $\mu \in \Delta_{\mathrm{F}}\mathfrak{A}$, $\{A_1, \ldots, A_L\} \subset \mathfrak{A}$ a partition of M, such that $\mu(A_h) > 0$ for $h = 1, \ldots, L$, and $(r_1, \ldots, r_L) \in \Delta_{\mathrm{F}}^L$. For

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

then

$$\pi_J(\mu)(A) = \sum_{h=1}^{L} r_h \mu(A \mid A_h) \qquad (A \in \mathfrak{A}). \tag{4.23}$$

Proof: J is as in theorem 4.0.19 with $J_h^* = \Delta_{\mathrm{F}}\mathfrak{A}^h$ for $h = 1, \ldots, L$. By (4.21) then

$$\nu^h = \mu^h$$

for all $\nu \in \Pi_J(\mu)$. Since ν is completely determined on $\bar{\mathfrak{A}}$ by J, this yields the unique solution (4.23) for $\nu \in \Pi_J(\mu)$. □

The foregoing lemmas and theorems show that what may be considered the fundamental properties of cross-entropy and cross-entropy minimization for probability measures on finite sets can be derived from the elementary properties FUN and BD of the logarithmic function. In addition to providing encouragement for considering cross-entropy minimization in the general context of lrc-fields, these results are also of some interest for the classical case of real-valued probabilities by showing how to derive properties like positivity and the reduction of cross-entropy minimization to Jeffrey's rule in a most elementary way from FUN and BD. Previous proofs of these properties (positivity (for measures with densities) in [Kullback and Leibler, 1951], Jeffrey's rule in [Diaconis and Zabell, 1982] (as part of a more general result) and [Lemmer and Barth, 1982] (for the finite case)) employed fairly powerful tools from analysis, and provided less insight into the basic analytic roots from which the behaviour of cross-entropy minimization derives.

Chapter 5

The Logic of Subjective Probabilities

5.1 Syntax

We define an extension of L^g_S for expressing subjective probabilities. The outward appearance of the extended syntax we define is again borrowed from [Bacchus, 1990b]. Since we are aiming at semantics for our language quite different from that supplied by Bacchus, the details of the definition are different, though.

In section 3.1 it has been observed that a subjective probability typically is a a number assigned to some uncertain event e and a property ϕ. A new term

$$\mathrm{prob}(\phi[\mathrm{e}]) \tag{5.1}$$

representing this number, therefore will be the basic building block for the extension of L^g_S.

Generally, we will want to make statements of degrees of belief not only for one single event e, but about several events $\mathrm{e}_1, \ldots, \mathrm{e}_n$. Such statements may also refer to the probability that two or more events are related in a certain way. The statement "with a probability greater than 0.6, this film is better than the one running on the other channel", for example, quantifies the relationship "better" between two uncertain objects.

This leads to the following precise definition for terms (5.1).

Definition 5.1.1 Let S be a vocabulary, $\mathbf{e}$ a tuple of *event symbols* not belonging to S, and $\phi(\boldsymbol{v}) \in \mathrm{L}^g_\mathrm{S}$ with $|\boldsymbol{v}|=|\mathbf{e}|$. Then

$$\mathrm{prob}(\phi[\boldsymbol{v}/\mathbf{e}]) \tag{5.2}$$

is called a *subjective probability* S*-term for* $\mathbf{e}$. Usually, we simply write $\mathrm{prob}(\phi[\mathbf{e}])$ for (5.2).

Note that in definition 5.1.1 it is required that for each free variable v in $\phi(\boldsymbol{v})$ an event symbol e is substituted, so that a subjective probability term does not contain free variables.

Apart from adding subjective probability terms, L^σ will also be extended by introducing the new function symbol ln. However, the use of ln will be more restricted than that of the function symbols $\cdot$ and $+$. This is already apparent from the definition of the subjective probability terms, where the condition $\phi(\boldsymbol{v}) \in \mathrm{L}^g_\mathrm{S}$ implies that ϕ may contain $\cdot$ and $+$, but not ln. An analogous restriction will be imposed on statistical probability terms in the new language.

Thus, we define the language $\mathrm{L}^{\beta}_{\mathrm{S},\mathbf{e}}$ (simply designated L^{β} when the reference to the specific sets of symbols can be dispensed with) by adopting the vocabulary $\mathrm{S}_{\mathrm{LOF}}$ of logarithmic ordered fields as a fixed part of the language, and inductively define the sets of terms and formulas of the language precisely as for L^{σ}(cf. p. 18), except that for field-terms rule (d) is slightly modified, and two further rules are added

(d′) If $\phi\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle$ is a formula not containing ln, then

$$[\phi\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle]_{\boldsymbol{w}}$$

is a field-term in the variables $\boldsymbol{v}$ and $\boldsymbol{x}$.

(e) If f is a field-terms, then $\mathsf{ln}(\mathrm{f})$ is a field-term.

(f) Every subjective probability S-term for $\mathbf{e}$ is a field-term.

The set of field-terms definable via (a)-(c),(d′),(e),and (f) is denoted $\mathrm{FT}^{\beta}_{\mathrm{S},\mathbf{e}}$.

Usually, we will have little use for the function symbol ln when expressing probabilistic information in L^{β}. Making it part of the language primarily serves the purpose to later be able to axiomatize default reasoning about probabilities within the language. Until that will be done in section 5.6, the symbol ln has to be taken into account when properties of L^{β} are shown, or definitions relating to L^{β} are given, but will not serve any practical purpose.

As for statistical probability terms, we introduce notation for conditional subjective probabilities: when t is a field-term containing the subterm

$$y \cdot \mathrm{prob}((\phi \wedge \psi)[\mathbf{e}]),$$

then the formula

$$\forall y(y \cdot \mathrm{prob}(\psi[\mathbf{e}]) = 1 \ \rightarrow \ \mathrm{t} \leq \mathrm{s})$$

is abbreviated by

$$\mathrm{t}' \leq \mathrm{s},$$

where t' is t with the term $y \cdot \mathrm{prob}((\phi \wedge \psi)[\mathbf{e}])$ replaced by the expression

$$\mathrm{prob}(\phi[\mathbf{e}] \mid \psi[\mathbf{e}]).$$

The examples from section 3.1 in this formal syntax are represented by

$$\begin{gathered}\mathrm{prob}(\mathsf{R}_1\mathrm{t} \wedge \mathsf{R}_2\mathrm{t} \wedge \mathsf{R}_3\mathrm{t}) = 0.3\\ \mathrm{prob}(\mathsf{A}\,\mathrm{f}) \leq 0.5\\ \mathrm{prob}(\mathsf{Actor}\,\mathsf{father}(\mathrm{p})) \geq 0.8.\end{gathered}$$

The first two of these representations are straightforward, using event symbols t (toss), f (film) and predicates R_i (result is i) and A (American). In the third example, the given representation is only one of several distinct possibilities. The least essential choice that has been made in this representation is to let ≥ 0.8 stand for "high probability". More fundamental is the decision to represent the given informal statement by a relation symbol Actor, a function symbol father, and an event symbol p (for the person referred to by "his"), and thereby to interpret the sentence as referring to an object with the uncertain property of having an actor as a father. Alternatively, one might have seen that father himself as an uncertain object f, leading to the representation

$$\mathrm{prob}(\mathsf{Actor}\,\mathrm{f}) \geq 0.8.$$

Also, the statement can be understood as referring to two persons p_1 and p_2, and an uncertain relation between them:

$$\mathrm{prob}(\mathsf{father}(\mathrm{p}_1) = \mathrm{p}_2 \wedge \mathsf{Actor}\,\mathrm{p}_2) \geq 0.8.$$

Similarly,

$$\mathrm{prob}(\mathsf{father}(\mathsf{h}) = \mathrm{p}_2 \wedge \mathsf{Actor}\,\mathrm{p}_2) \geq 0.8,$$

but here the unnamed subject referred to by "his" is represented by a standard constant symbol h, and only his father considered to be an uncertain object.

By this example it can be seen that in addition to the usual array of choices for how to encode a natural language sentence in formal logic, here we also have to decide which of the objects referred to in a statement are meant to account for the uncertainty expressed, and therefore should be represented by event symbols, and which can be understood as fully determined objects best represented by constant symbols.

Some further examples of L^β-formulas are:

"This film has less than average probability of being American":

$$\mathrm{prob}(\mathsf{A}\,\mathrm{f}) \leq [\mathsf{A}v]_v.$$

"The probability that this film has a happy end is the probability that it is American multiplied by the statistical probability for happy endings in American films, plus the probability that it is not American multiplied by the statistical probability for happy endings in non-American films." (Jeffrey's rule):

$$\mathrm{prob}(\mathsf{HE}\,\mathrm{f}) = \mathrm{prob}(\mathsf{A}\,\mathrm{f}) \cdot [\mathsf{HE}v \mid \mathsf{A}v]_v + \mathrm{prob}(\neg\mathsf{A}\,\mathrm{f}) \cdot [\mathsf{HE}v \mid \neg\mathsf{A}v]_v.$$

"This film is an American production, and it is very likely to have a happy end"

$$\mathrm{prob}(\mathsf{A}\,\mathrm{f}) = 1 \wedge \mathrm{prob}(\mathsf{HE}\,\mathrm{f}) \geq 0.8.$$

L^β does not allow any free use of the event symbols outside a prob()-operator, so that definite statements about these events can only be approximated by stating a probability of 1.

L^β has an appearance very similar to the languages defined by Halpern [1990] and Bacchus [1990b], both of which, following Halpern, we refer to as $\mathscr{L}_3^=$. The most basic distinguishing feature of the two languages, of course, is the use of a special set $\mathbf{e}$ of new symbols in $\mathrm{L}^\beta_{\mathrm{S},\mathbf{e}}$ that are treated differently from the standard constant symbols in S. In $\mathscr{L}_3^=$, formulas inside the probability operator prob() are allowed to contain free variables, as well as the operator prob() again. Neither is possible in L^β.

The second of these two limitations appears to be rather less serious because "meta-degrees of belief", i.e. degrees of belief about one's own degrees of belief are a fairly unnatural concept. Free variables within prob(), on the other hand, are a useful tool to represent more complex statements about subjective probabilities: "Only few films ($\leq 10\%$) are very likely to be better than f", for instance, can be represented in $\mathscr{L}_3^{=}$, but not in L^β, by

$$[\mathrm{prob}(\mathsf{Better}\, v\mathrm{f}) \geq 0.9]_v \leq 0.1$$

using a free variable v inside prob(). Compare this statement to "It is very likely that only few films are better than f", which is representable in L^β by

$$\mathrm{prob}([\mathsf{Better}\, v\mathrm{f}]_v \leq 0.1) \geq 0.9.$$

One reason for imposing the restriction of applying prob() only to closed L^σ-formulas in L^β is that this will simplify the semantical definitions in the following section. While it would be possible to use the general semantic principle there introduced also for interpreting a language permitting arbitrary L^β-formulas inside prob(), we will not follow such a general approach here.

The reason for this is that more than in sound semantics for a very expressive language – as is aptly supplied by Bacchus and Halpern – we are here interested in strengthening such a semantics to incorporate default reasoning about probabilities. It will turn out, that the definition of this strengthened semantics in section 5.4 makes essential use of the fact that subjective probability terms do not contain free variables.

5.2 Semantics: Feasible Models

Viewing statements of subjective probabilities as statements about a particular set of uncertain objects or events is the key to using a semantic construction for the interpretation of these statements that is instrumental for our approach to combining statistical knowledge with subjective beliefs.

A statement of a subjective probability r that an event e has some property ϕ is equivalent to saying that with probability r, e belongs to the set of all events that have property ϕ. Note that this equivalence, obvious though it may appear, is not completely trivial, because it reduces the quantified relation between e and the abstract syntactic construct ϕ to a relation between e and the extension of ϕ on the object level. By virtue of this equivalence it is then straightforward to view uncertain events as random elements of the domain of discourse, about which some partial knowledge has been obtained. They are then naturally interpreted by a probability measure on the domain of a statistical S-structure. This approach to interpreting subjective probabilities we call *random event semantics*.

With the syntax introduced in section 5.1, subjective probabilities are expressible only for subsets of $M^{|\mathbf{e}|}$ that are definable in L^σ without parameters – a consequence of not allowing free variables inside prob$(\cdot)$. For the interpretation of subjective probability terms therefore a probability measure on the subalgebra of $\mathfrak{A}_{|\mathbf{e}|}$ consisting of such sets will be sufficient.

Definition 5.2.1 Let S be a vocabulary, $\mathbf{e}$ a tuple of event symbols. $(\mathfrak{M}, \nu_{\mathbf{e}})$ is a *belief* S-*structure* for $\mathbf{e}$, if

- $\mathfrak{M} = (M, I, \mathfrak{F}, (\mathfrak{A}_n, \mu_n)_n)$, where $\mathfrak{F}$ is an lrc-field, and $\mathfrak{M}\restriction \mathrm{S_{OF}} = (M, I, \mathfrak{F}\restriction \mathrm{S_{OF}}, (\mathfrak{A}_n, \mu_n)_n)$ is a statistical S-structure.

- $\nu_{\mathbf{e}}$ is an $\mathbf{F}$-valued probability measure on

$$\mathfrak{A}_{\mathbf{e}} := \{\mathfrak{M}(\phi\langle \boldsymbol{v}\rangle) \mid |\boldsymbol{v}|=|\mathbf{e}|,\ \phi \in \mathrm{L}_{\mathrm{S}}^{\sigma}\} \subseteq \mathfrak{A}_{|\mathbf{e}|}.$$

$\mathfrak{M}$ is called the *statistical base structure* (of the belief structure $(\mathfrak{M},\nu_{\mathbf{e}})$). When $\mathfrak{F} = \mathfrak{R}$, $(\mathfrak{M},\nu_{\mathbf{e}})$ is called a *real-valued belief structure*

It is a little bit unfortunate that we here have to distinguish between statistical structures, as introduced in definition 2.3.4, on the one hand, and statistical base structures on the other. The only difference being that the real closed field in the latter must be equipped with a logarithmic function, which was not required for the former. This terminological awkwardness might have been avoided, if lrc-fields had been required as the range for probability values right from the beginning. Since this would have served no discernible purpose for the interpretation of L^{σ}, however, it is perhaps preferable to here suffer the distinction between statistical structures and statistical base structures.

The satisfaction relation for statistical S-structures $\mathfrak{M}$ and $\mathrm{L}_{\mathrm{S}}^{\sigma}$-formulas is easily extended to a satisfaction relation between belief S-structures $(\mathfrak{M},\nu_{\mathbf{e}})$ and $\mathrm{L}_{\mathrm{S},\mathbf{e}}^{\beta}$-formulas: let γ be a variable assignment. For $((\mathfrak{M}, \nu_{\mathbf{e}}), \gamma)$, terms and formulas in $\mathrm{L}_{\mathrm{S},\mathbf{e}}^{\beta}$, a relation $\models_{\beta}$ is defined precisely as the relation $\models_{\sigma}$ in definition 2.3.1 by replacing $(\mathfrak{M}, \gamma)$ with $((\mathfrak{M}, \nu_{\mathbf{e}}), \gamma)$, and $\models_{\sigma}$ with $\models_{\beta}$ throughout, and adding the new conditions

(e) $\mathrm{t} \equiv \mathsf{ln}(\mathrm{t}')$. Then $((\mathfrak{M}, \nu_{\mathbf{e}}), \gamma)(\mathrm{t}) = \mathsf{ln}^{\mathrm{F}}(((\mathfrak{M}, \nu_{\mathbf{e}}), \gamma)(\mathrm{t}'))$.

(f) $\mathrm{t} \equiv \mathrm{prob}(\phi[\boldsymbol{v}/\mathbf{e}])$. Then $((\mathfrak{M}, \nu_{\mathbf{e}}), \gamma)(\mathrm{t}) = \nu_{\mathbf{e}}(\{\boldsymbol{a} \mid ((\mathfrak{M}, \nu_{\mathbf{e}}), \gamma[\boldsymbol{v}/\boldsymbol{a}]) \models_{\beta} \phi(\boldsymbol{v})\})$.

Since $\phi(\boldsymbol{v})$ here is a formula in L^{σ}, $\mathfrak{M}\restriction \mathrm{S_{OF}}$ was assumed to be a statistical S-structure, and for $\phi \in \mathrm{L}^{\sigma}$

$$((\mathfrak{M}, \nu_{\mathbf{e}}), \gamma) \models_{\beta} \phi(\boldsymbol{v}) \quad \Leftrightarrow \quad (\mathfrak{M}\restriction \mathrm{S_{OF}}, \gamma) \models_{\sigma} \phi(\boldsymbol{v}), \tag{5.3}$$

it is assured that the set

$$\{\boldsymbol{a} \mid ((\mathfrak{M}, \nu_{\mathbf{e}}), \gamma[\boldsymbol{v}/\boldsymbol{a}]) \models_{\beta} \phi(\boldsymbol{v})\}$$

belongs to $\mathfrak{A}_{\mathbf{e}}$ (to match the formal definition of $\mathfrak{A}_{\mathbf{e}}$, write this set as $\mathfrak{M}(\hat{\phi}\langle \boldsymbol{v}\rangle)$, where $\hat{\phi}\langle \boldsymbol{v}\rangle$ is the conjunction of $\phi(\boldsymbol{v})$ and the atomic formulas $v = v$ for all $v \in \boldsymbol{v}$ that do not actually appear in $\phi(\boldsymbol{v})$ (cf. p. 25)).

Furthermore, by the restriction of applying the prob()-operator only to closed formulas, the subjective probability terms can not be used to define subsets of M^n not definable in $\mathrm{L}_{\mathrm{S}}^{\sigma}$ alone. Hence, also for every statistical probability term $\mathrm{t} \equiv [\phi(\boldsymbol{v},\boldsymbol{w},\boldsymbol{x})]_{\boldsymbol{w}} \in \mathrm{FT}_{\mathrm{S},\mathbf{e}}^{\beta}$, the interpretation $((\mathfrak{M}, \nu_{\mathbf{e}}), \gamma)(\mathrm{t})$ is defined. This is formally stated in the following lemma.

Lemma 5.2.2 Let $(\mathfrak{M},\nu_{\mathbf{e}})$ be a belief S-structure, γ a variable assignment. Then $((\mathfrak{M}, \nu_{\mathbf{e}}), \gamma)(\mathrm{t})$ is defined for every $\mathrm{t} \in \mathrm{FT}_{\mathrm{S},\mathbf{e}}^{\beta}$.

Proof: It must be shown that for all $\phi\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle \in \mathrm{L}^{\beta}_{\mathrm{S},\mathbf{e}}$ not containing ln, $\boldsymbol{a} \in M^{|\boldsymbol{v}|}$, and $\boldsymbol{r} \in M^{|\boldsymbol{x}|}$, the set $((\mathfrak{M}, \nu_{\mathbf{e}}), \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\phi\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle)$ is measurable. The interpretation

$$s_i := ((\mathfrak{M}, \nu_{\mathbf{e}}), \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\mathrm{t}_i)$$

of every subjective probability term $\mathrm{t}_i \equiv \mathrm{prob}(\psi_i[\mathbf{e}])$ appearing in ϕ is a constant not depending on the variable assignment $\gamma(\boldsymbol{v}) = \boldsymbol{a}, \gamma(\boldsymbol{x}) = \boldsymbol{r}$. Define $\phi'\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x},\boldsymbol{y}\rangle \in \mathrm{L}^{\sigma}_{\mathrm{S}}$ by replacing the t_i with new field variables y_i. Then

$$\begin{aligned}&((\mathfrak{M}, \nu_{\mathbf{e}}), \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r})(\phi\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle) = \\ &(\mathfrak{M}\restriction \mathrm{S}_{\mathrm{OF}}, \boldsymbol{v}/\boldsymbol{a}, \boldsymbol{x}/\boldsymbol{r}, \boldsymbol{y}/\boldsymbol{s})(\phi'\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x},\boldsymbol{y}\rangle) \in \mathfrak{A}_{|\boldsymbol{w}|}.\end{aligned}$$

□

The interpretation of $\mathrm{prob}(\phi[\boldsymbol{v}/\mathbf{e}])$ depends on the order imposed on the event symbols by writing them as a tuple $\mathbf{e}$ because without this order the set whose measure is the interpretation of $\mathrm{prob}(\phi[\boldsymbol{v}/\mathbf{e}])$ would only be determined up to permutations. Since $\nu_{\mathbf{e}}$ is not assumed to be homogeneous, this is not sufficient.

The ability to obtain structures for the interpretation of $\mathrm{L}^{\beta}_{\mathrm{S},\mathbf{e}}$ simply by adding to some existing statistical S-structure the definition of a logarithmic function (when this is possible in $\mathfrak{F}$) and a measure $\nu_{\mathbf{e}}$ on $\mathfrak{A}_{\mathbf{e}}$ is one reason for restricting the syntax of probability terms. If subjective probability terms were allowed to contain free variables, then definability in $\mathrm{L}^{\beta}_{\mathrm{S},\mathbf{e}}$ would be more powerful than in $\mathrm{L}^{\sigma}_{\mathrm{S}}$, making the algebraic structure $(\mathfrak{A}_n)$ in $\mathfrak{M}$ insufficient for interpreting $\mathrm{L}^{\beta}_{\mathrm{S},\mathbf{e}}$. The same would be true if the symbol ln was allowed inside $[\cdot]$ or $\mathrm{prob}(\cdot)$.

Definition 5.2.3 A belief S-structure $(\mathfrak{M},\nu_{\mathbf{e}})$ for $\mathbf{e}$ is called a *feasible model* of a set Φ of $\mathrm{L}^{\beta}_{\mathrm{S},\mathbf{e}}$-sentences (written $(\mathfrak{M}, \nu_{\mathbf{e}}) \models_{\beta} \Phi$), if $(\mathfrak{M}, \nu_{\mathbf{e}}) \models_{\beta} \phi$ for every $\phi \in \Phi$.

Definition 5.2.4 Let $\Phi \subseteq \mathrm{L}^{\beta}_{\mathrm{S},\mathbf{e}}$, $\phi \in \mathrm{L}^{\beta}_{\mathrm{S},\mathbf{e}}$. ϕ is β-entailed by Φ, written $\Phi \models_{\beta} \phi$, if every belief S-structure that is a feasible model of Φ, also is a feasible model of ϕ. Also, we use the notation $\Phi \models^{\mathbf{R}}_{\beta} \phi$ if every real-valued feasible model of Φ is a feasible model of ϕ.

Definition 5.2.5 We write $\mathscr{L}^{\beta}$ for the logic defined by the language L^{β} and the entailment relation $\models_{\beta}$.

Similarly as for $\models_{\sigma}$, we have that $\models_{\beta}$ coincides with standard first-order entailment for first-order formulas Φ, ψ:

$$\Phi \models_{\beta} \psi \;\Leftrightarrow\; \Phi \models \psi \qquad\qquad (\Phi, \psi \text{ first-order})$$

(cf. lemma 2.3.13). By the given definitions it is not assured, however, that $\models_{\beta}$ coincides with $\models_{\sigma}$ on L^{σ}-formulas. Since the relation $\models_{\beta}$ is based on the more restrictive concept of a model in which measures are taken in an lrc-field, not merely an rc-field, we immediately only obtain that for $\Phi \subseteq \mathrm{L}^{\sigma}, \psi \in \mathrm{L}^{\sigma}$

$$\Phi \models_{\sigma} \phi \;\Rightarrow\; \Phi \models_{\beta} \phi. \tag{5.4}$$

For the $\models^{\mathbf{R}}_\sigma$ and $\models^{\mathbf{R}}_\beta$ - relations, on the other hand, we clearly have

$$\Phi \models^{\mathbf{R}}_\sigma \psi \;\Leftrightarrow\; \Phi \models^{\mathbf{R}}_\beta \psi$$

for $\Phi \subseteq \mathrm{L}^\sigma, \psi \in \mathrm{L}^\sigma$. An incompleteness result immediately follows for $\models^{\mathbf{R}}_\beta$ from the incompleteness of $\models^{\mathbf{R}}_\sigma$.

A sufficient condition for the converse of (5.4) to be true would be that for every statistical structure $\mathfrak{M} = (M, \ldots, \mathfrak{F}, \ldots)$ there exists a statistical structure $\mathfrak{M}^* = (M, \ldots, \mathfrak{F}^*, \ldots)$ where $\mathfrak{F}^*$ is an rc-field that can be extended by a logarithmic function, and

$$\forall \phi \in \mathrm{L}^\sigma \quad \mathfrak{M} \models_\sigma \phi \;\Leftrightarrow\; \mathfrak{M}^* \models_\sigma \phi. \tag{5.5}$$

While it may be conjectured that such $\mathfrak{M}^*$ indeed always exist, and hence an equivalence in (5.4) actually holds, this issue shall not be pursued here further, so that, for the time being, we have to distinguish the two entailment relations $\models_\sigma$ and $\models_\beta$ on L^σ.

5.3 $\mathscr{L}^\beta$ Is First-Order Logic

For $\mathscr{L}^\beta$ we have essentially the same reduction to first-order logic as for $\mathscr{L}^\sigma$. We here sketch out the derivation of this result without spelling out all the details, which are the same as in the corresponding results from section 2.5.

Let S be a vocabulary and $\mathbf{e}$ a tuple of event symbols. The translation of $\mathrm{L}^\sigma_\mathrm{S}$ into a first-order language $\mathrm{L}_{\mathrm{S}_\infty}$ defined in section 2.5.2 can be extended to a translation of $\mathrm{L}^\beta_{\mathrm{S},\mathbf{e}}$ into a first-order language $\mathrm{L}_{\mathrm{S}'_\infty}$ with $\mathrm{S}'_\infty \supset \mathrm{S}_\infty$. The additional rules needed for translating subjective probability terms $\mathrm{prob}(\phi[\boldsymbol{v}/\mathbf{e}])$ are much simpler than what was needed for statistical probability terms $[\phi(\boldsymbol{v},\boldsymbol{w},\boldsymbol{x})]_{\boldsymbol{w}}$ because the former do not contain free variables, and therefore behave as constants, not as functions.

The vocabulary S'_∞ is defined similarly as S_∞, introducing additional constant symbols to be used as the translation of subjective probability terms. We will here make the same assumptions about the set of domain- and field variables used being $\{v_1, v_2, \ldots\}$ and $\{x_1, x_2, \ldots\}$ respectively. Besides $\{\alpha_1, \alpha_2, \ldots\}$ and $\{\zeta_1, \zeta_2, \ldots\}$, we use a further (finite) set $\{\epsilon_1, \ldots, \epsilon_{|\mathbf{e}|}\}$ of auxiliary variables. The tuple $(\epsilon_1, \ldots, \epsilon_{|\mathbf{e}|})$ is denoted by $\boldsymbol{\epsilon}$.

Let $\mathrm{S}_0, \mathrm{S}_1, \ldots, \mathrm{S}_\infty$ be defined as in section 2.5.2. For each n a vocabulary $\mathrm{S}'_n \supset \mathrm{S}_n$ is defined. For $n = 0$ let

$$\mathrm{S}'_0 := \mathrm{S} \cup \{0, 1, +, \cdot, \leq, \mathsf{ln}\} \supset \mathrm{S}_0.$$

Now let $n \geq 0$, and assume that $\mathrm{S}'_n \supset \mathrm{S}_n$ has been defined. The pattern $\mathrm{p}(\phi; \boldsymbol{v})$ of $\phi\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle \in \mathrm{L}_{\mathrm{S}'_n}$ with respect to $\boldsymbol{v}$ is defined as in section 2.5.2. For each formula

$$\phi\langle \boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle \in \mathrm{L}_{\mathrm{S}'_n \setminus \{\mathsf{ln}\}} \setminus \mathrm{L}_{\mathrm{S}'_{n-1}}$$

and permutation $\pi\boldsymbol{w}$ of $\boldsymbol{w}$ let

$$\mathsf{f}^{\mathrm{p}(\phi;\boldsymbol{v}),\pi\boldsymbol{w}}$$

be a new function symbol of sort $\mathrm{D}^{|\boldsymbol{v}|}\mathrm{F}^{|\boldsymbol{x}|+1}$. For each

$$\phi(\boldsymbol{\epsilon}) \in \mathrm{L}_{\mathrm{S}_n}$$

let

$$\mathsf{q}^{\phi(\boldsymbol{\epsilon})}$$

be a new constant symbol of sort F. Observe that these constant symbols are only generated from $\phi \in \mathrm{L}_{\mathrm{S}_n}$, not from $\phi \in \mathrm{L}_{\mathrm{S}'_n}$. Let S'_{n+1} denote the union of S'_n and these two sets of new symbols. Clearly $\mathrm{S}'_{n+1} \supset \mathrm{S}_{n+1}$, because $\mathrm{S}_{n+1} \supset \mathrm{S}_n$ consists just of those symbols $\mathsf{f}^{\mathrm{p}(\phi;\boldsymbol{v}),\pi\boldsymbol{w}}$ with $\phi \in \mathrm{L}_{\mathrm{S}_n} \subset \mathrm{L}_{\mathrm{S}'_n \setminus \{\mathsf{ln}\}}$. Define

$$\mathrm{S}'_\infty := \bigcup_{n\in\mathbf{N}} \mathrm{S}'_n, \qquad \mathrm{S}'_+ := \mathrm{S}'_\infty \setminus \mathrm{S}'_0.$$

The translations

$$\mathrm{t} \mapsto \mathrm{t}^* \quad \text{and} \quad \phi \mapsto \phi^*$$

for $\mathrm{L}^\beta_{\mathrm{S},\mathbf{e}}$- terms and formulas is defined as for L^σ(p. 48), with the additional rules for field terms

(e) $(\mathsf{ln}(\mathrm{t}))^* :\equiv \mathsf{ln}(\mathrm{t}^*)$

(f) $(\mathrm{prob}(\phi[\boldsymbol{v}/\mathbf{e}]))^* :\equiv \mathsf{q}^{\phi^*[\boldsymbol{v}/\boldsymbol{\epsilon}]}$.

Note that the augmented set of translation rules will always yield $\phi^* \in \mathrm{L}_{\mathrm{S}_\infty}$ when $\phi \in \mathrm{L}^\sigma$ (as required within the operator prob()), so that the symbol $\mathsf{q}^{\phi^*[\boldsymbol{v}/\boldsymbol{\epsilon}]}$ actually exists. Similarly, for the old rule

$$([\phi\langle\boldsymbol{v},\boldsymbol{w},\boldsymbol{x}\rangle]_{\boldsymbol{w}})^* :\equiv \mathsf{f}^{\mathrm{p}(\phi^*;\boldsymbol{v}),\boldsymbol{w}}(\sigma\boldsymbol{v},\sigma\boldsymbol{x})$$

it must be observed that the translation ϕ^* does not contain the symbol ln when ϕ does not, so that the symbol $\mathsf{f}^{\mathrm{p}(\phi^*;\boldsymbol{v}),\boldsymbol{w}}$ exists.

An inverse mapping $\mathrm{L}_{\mathrm{S}'_\infty} \to \mathrm{L}^\beta_{\mathrm{S},\mathbf{e}}$ can not be defined as a total function as was done for the mapping $\mathrm{L}_{\mathrm{S}_\infty} \to \mathrm{L}^\sigma_{\mathrm{S}}$: since in $\mathrm{L}^\beta_{\mathrm{S},\mathbf{e}}$ statistical probability terms $[\phi(\boldsymbol{v},\boldsymbol{w},\boldsymbol{x})]_{\boldsymbol{w}}$ are not allowed to contain the symbol ln, the inverse can not be (easily) explained for $\mathrm{L}_{\mathrm{S}'_\infty}$-terms of the form

$$\mathsf{f}^{\mathrm{p},\boldsymbol{w}}(\ldots,\mathsf{ln}(\ldots),\ldots) \qquad (\mathsf{f}^{\mathrm{p},\boldsymbol{w}} \in \mathrm{S}'_+).$$

Hence we restrict the definition of $(\cdot)^{-1}$ to the fragment $\tilde{\mathrm{L}}_{\mathrm{S}'_\infty}$ of $\mathrm{L}_{\mathrm{S}'_\infty}$ defined by the restriction that in a term of the form $\mathsf{f}^{\mathrm{p},\boldsymbol{w}}(\mathrm{t}_1,\ldots,\mathrm{t}_k,\mathrm{s}_1,\ldots,\mathrm{s}_l)$ none of the field-terms $\mathrm{s}_1,\ldots,\mathrm{s}_l$ may contain a subterm of the form $\mathsf{ln}(\mathrm{s})$.

For $\mathrm{t},\phi \in \tilde{\mathrm{L}}_{\mathrm{S}'_\infty}$ then t^{-1} and ϕ^{-1} are defined as in section 2.5.2, with the additional rules

- $(\mathsf{ln}(\mathrm{s}))^{-1} :\equiv \mathsf{ln}(\mathrm{s}^{-1})$.
- Let $\mathsf{q}^{\phi(\boldsymbol{\epsilon})} \in \mathrm{S}'_n$. Assume that $(\phi(\boldsymbol{\epsilon}))^{-1} \in \mathrm{L}^\sigma_{\mathrm{S}}$ has been defined. Then

$$(\mathsf{q}^{\phi(\boldsymbol{\epsilon})})^{-1} :\equiv \mathrm{prob}((\phi(\boldsymbol{\epsilon}))^{-1}[\boldsymbol{\epsilon}/\mathbf{e}]).$$

Clearly t^{-1} then does not contain ln when t does not contain ln, so that by the definition of $(\mathsf{f}^{\mathrm{p},\boldsymbol{w}}(\ldots))^{-1}$ (cf. p. 50) ln will not be introduced into a statistical probability term $[\ldots]$.

Lemmas 2.5.7 and 2.5.8 remain valid for the generalized translations $(\cdot)^*$ and $(\cdot)^{-1}$. The proofs are as before; for lemma 2.5.8 it must now also be noted that $\phi^* \in \tilde{\mathrm{L}}_{\mathrm{S}'_\infty}$ for $\phi \in \mathrm{L}^\beta_{\mathrm{S},\mathbf{e}}$, so that $(\phi^*)^{-1}$ is defined.

An $\mathrm{L}_{\mathrm{S}'_\infty}$-structure corresponding to a belief structure $(\mathfrak{M},\nu_{\mathbf{e}})$ is defined as in section 2.5.3. The interpretation of a constant symbol $\mathsf{q}^{\phi(\boldsymbol{\epsilon})} \in \mathrm{S}'_\infty$ is given by

$$I^*(\mathsf{q}^{\phi(\boldsymbol{\epsilon})}) := (\mathfrak{M}, \nu_{\mathbf{e}})(\mathrm{prob}(\phi[\mathbf{e}])).$$

Lemma 2.5.9 then holds for $\tilde{\mathrm{L}}_{\mathrm{S}'_\infty}$- terms, formulas and sentences.

An axiomatization $\mathrm{AX}(\mathbf{e})$ of the class of S'_∞-structures corresponding to belief S-structures for $\mathbf{e}$ is given by a modification of the axiom set AX (p. 54). First, in $\mathrm{AX}(\mathbf{e})$, the axioms LRCF take the place of RCF. The schemas (2.25),(2.28), and (2.30) that in AX are instantiated for all $\mathrm{L}_{\mathrm{S}_\infty}$- formulas, in $\mathrm{AX}(\mathbf{e})$ are instantiated by all $\tilde{\mathrm{L}}_{\mathrm{S}'_\infty}$-formulas. The definition of $\mathrm{AX}(\mathbf{e})$ is completed by adding axioms that make sure that the interpretations of the $\mathsf{q}^{\phi(\boldsymbol{\epsilon})}$ encode a probability measure on the $\mathrm{L}_{\mathrm{S}_\infty}$-definable subsets of $M^{|\mathbf{e}|}$:

- Let $\mathrm{PM}(\mathbf{e})$ contain
  - For all constant symbols $\mathsf{q}^{\phi(\boldsymbol{\epsilon})} \in \mathrm{S}'_\infty$ the axiom
  $$\mathsf{q}^{\phi(\boldsymbol{\epsilon})} \geq 0.$$
  - With $\tau(\boldsymbol{\epsilon}) :\equiv (\epsilon_1 = \epsilon_1 \wedge \ldots \wedge \epsilon_{|\mathbf{e}|} = \epsilon_{|\mathbf{e}|})$ the axiom
  $$\mathsf{q}^{\tau(\boldsymbol{\epsilon})} = 1.$$
  - For all $\phi(\boldsymbol{\epsilon}), \psi(\boldsymbol{\epsilon}) \in \mathrm{L}_{\mathrm{S}_\infty}$ the axiom
  $$\neg\exists\boldsymbol{\epsilon}\,(\phi(\boldsymbol{\epsilon}) \wedge \psi(\boldsymbol{\epsilon})) \rightarrow \mathsf{q}^{(\phi\vee\psi)(\boldsymbol{\epsilon})} = \mathsf{q}^{\phi(\boldsymbol{\epsilon})} + \mathsf{q}^{\psi(\boldsymbol{\epsilon})}.$$

The analogue of lemma 2.5.10 now holds for belief structures $(\mathfrak{M},\nu_{\mathbf{e}})$ and $\mathrm{AX}(\mathbf{e})$.

The last step in reducing $\mathscr{L}^\beta$ to first-order logic now is the definition of a belief structure $(\mathfrak{M}^{-1}, \nu_{\mathbf{e}})$ from an S'_∞-structure $\mathfrak{M} \models \mathrm{AX}(\mathbf{e})$. This is done similarly as for S_∞-structures in section 2.5.3. The algebras $\mathfrak{A}_n$ in $\mathfrak{M}^{-1}$ are the sets of all $\tilde{\mathrm{L}}_{\mathrm{S}'_\infty}$-definable subsets of $M^{|\mathbf{e}|}$. (As a matter of fact, these are just the same as the $\mathrm{L}_{\mathrm{S}'_\infty}$-definable sets, because for every formula $\phi(\boldsymbol{v}) \in \mathrm{L}_{\mathrm{S}'_\infty}$ there exists an equivalent "term-reduced" formula $\phi_0(\boldsymbol{v}) \in \mathrm{L}_{\mathrm{S}'_\infty}$ in which function terms only have variables as arguments (see e.g. [Ebbinghaus *et al.*, 1984]). Such ϕ_0, in particular, belong to the fragment $\tilde{\mathrm{L}}_{\mathrm{S}'_\infty}$.)

The statistical measures (μ_n) are defined as before by (2.31). The belief measure $\nu_{\mathbf{e}}$ is obtained from the interpretations of the constant symbols q in $\mathfrak{M}$ via

$$\nu_{\mathbf{e}}(\mathfrak{M}^{-1}(\phi\langle\boldsymbol{v}\rangle)) := I(\mathsf{q}^{\phi^*[\boldsymbol{v}/\boldsymbol{\epsilon}]}) \qquad (\phi\langle\boldsymbol{v}\rangle \in \mathrm{L}^\sigma_{\mathrm{S}}).$$

From $\mathrm{PM}(\mathbf{e})$ it follows that this defines a probability measure on $\mathfrak{A}_{\mathbf{e}}$.

The analogue of lemma 2.5.11 then can be formulated for S'_∞-structures $\mathfrak{M}$ and the inverse $(\mathfrak{M}^{-1}, \nu_{\mathbf{e}})$. To part (b) of the original proof of lemma 2.5.11 a trivial induction step for field-terms of the form $\mathrm{prob}(\phi[\mathbf{e}])$ has to be added.

Piecing things together, we obtain the generalization of theorem 2.5.12.

Theorem 5.3.1 For all $\Phi \subset \mathrm{L}^{\beta}_{\mathrm{S},\mathbf{e}}$, $\phi \in \mathrm{L}^{\beta}_{\mathrm{S},\mathbf{e}}$:

$$\Phi \models_{\beta} \phi \ \text{ iff } \ \Phi^{*} \cup \mathrm{AX}(\mathbf{e}) \models \phi^{*}.$$

5.4 Semantics: Default Models

The concept of a feasible model for an L^{β}-formula is just as far as probability theory alone will take us. That is to say: semantics that allows any feasible model as a model for an $\mathrm{L}^{\beta}_{\mathrm{S},\mathbf{e}}$-theory Φ will permit just those inferences that are validated by probability calculus alone. From the perspective of probability theory, the information on the statistical measures (μ_n) contained in Φ, and the constraints on the belief measure $\nu_{\mathbf{e}}$, can be satisfied independently; any set of measures $\{(\mu_n)_n, \nu_{\mathbf{e}}\}$ that satisfies the respective constraints is as good as any other set.

In order to also incorporate some of the mechanisms of default reasoning about probabilities in our semantics, by which a rational agent will select from the multitude of feasible models those that appear more plausible than others, we will have to require something more of a belief structure for being a model of a theory, than mere feasibility in the sense of definition 5.2.3.

This strengthening of the notion of a feasible model for $\Phi \subseteq \mathrm{L}^{\beta}$ for formalizing default reasoning about probabilities will only be defined for finite theories Φ, or, equivalently (by identifying a finite set Φ with $\phi :\equiv \wedge\Phi$), single sentences $\phi \in \mathrm{L}^{\beta}$.

Generally speaking, what will be proposed here, is a *preferred model semantics* for $\mathrm{L}^{\beta}_{\mathrm{S},\mathbf{e}}$: a certain subclass of feasible models of $\phi \in \mathrm{L}^{\beta}_{\mathrm{S},\mathbf{e}}$ will be regarded as capturing the default meaning of ϕ, while others are rejected for this purpose. This makes the formalization of default reasoning about probabilities here developed similar to logical default reasoning systems based on preferred model semantics, notably circumscription ([McCarthy, 1980]).

Intuitively speaking, the criterion for either accepting or rejecting $(\mathfrak{M},\, \nu_{\mathbf{e}})$ as a preferred model of ϕ, or, as we shall say, a *default model* of ϕ, should be a restriction formulated for $\nu_{\mathbf{e}}$, rather than $\mathfrak{M}$. Our intention of not making any default inferences of objective statements implies that a feasible model $(\mathfrak{M},\nu_{\mathbf{e}})$ will not be discarded for the reason of the structure of $\mathfrak{M}$ alone, because all models for the objective information are considered equally valid. The criterion we look for, then will have to be of the form: $(\mathfrak{M},\nu_{\mathbf{e}})$ is a default model, if $\nu_{\mathbf{e}}$ is the preferred belief measure, given $\mathfrak{M}$ as the interpretation of objective sentences. More formally, let

$$\Delta_{\mathrm{F}}(\phi, \mathfrak{M}) \subseteq \Delta_{\mathrm{F}}\mathfrak{A}_{\mathbf{e}}$$

designate the set of belief measures $\nu_{\mathbf{e}}$ for which $(\mathfrak{M},\nu_{\mathbf{e}})$ is a feasible model of ϕ. We will specify a rule

$$\phi, \mathfrak{M} \mapsto \mathit{pref} \subseteq \Delta_{\mathrm{F}}(\phi, \mathfrak{M})$$

that assigns to the statistical structure $\mathfrak{M}$, and the $\mathrm{L}^{\beta}_{\mathrm{S},\mathbf{e}}$-formula ϕ a subset *pref* of preferred belief measures, so that $(\mathfrak{M},\nu_{\mathbf{e}})$ will be regarded a default model iff $\nu_{\mathbf{e}} \in \mathit{pref}$.

From our analysis of default reasoning about probabilities in section 3.3 and 3.4, it is clear that cross-entropy minimization will have to be the central tool to define preferred belief measures. There are still some questions to be answered, however, about how the minimum cross-entropy principle as there derived can be adjusted to the situation at hand: in our epistemic analysis we were only concerned with a single event e; how do we generalize these considerations to L^{β}-theories about tuples $\mathbf{e}$ of events, perhaps expressing degrees of belief about how the elements of $\mathbf{e}$ are related? Secondly, our derivation of the minimum cross-entropy principle in section 3.4 assumed a finitary context, i.e. events were modeled by random variables with a finite range; is this model sufficient to derive semantics for L^{β} where there are infinitely many properties definable for $\mathbf{e}$?

The problem of how to extend the model developed in chapter 3 for dealing with multiple events $\mathbf{e}$ rather than a single event e, is not difficult to resolve. There is no inherent difference between a set $\mathbf{e}$ of events, possibly related in certain ways, and a single event e with various properties: $\mathbf{e}$ can just be seen as a single composite event for which properties may be defined in terms of relations among its components. Thus, everything that has been said in section 3.2 about a single event e, without modifications, carries over to a tuple $\mathbf{e}$.

The formalization of subjective probability statements by L^{β}- sentences which are interpreted by belief structures, however, entails certain restrictions for what kinds of composite events can be modeled. Since in this formal framework, the domain of possible (composite) events is of the form $M^{|\mathbf{e}|}$, every component of $\mathbf{e}$ is assumed to stem from the same set M of possible (simple) events. Furthermore, in a belief structure, the algebra $\mathfrak{A}_{\mathbf{e}}$ always is equipped with a statistical measure

$$\mu_{\mathbf{e}} := \mu_{|\mathbf{e}|} \upharpoonright \mathfrak{A}_{\mathbf{e}}$$

that has the product property. Thus, making the assumption of postulate 2, here also necessarily entails an independence assumption for the components of $\mathbf{e}$: assuming the random mechanism that produced $\mathbf{e}$ to be equivalent to $\mu_{\mathbf{e}}$, particularly means that, without any evidence, one would for each pair $\mathbf{e}(i)$, $\mathbf{e}(j)$ of components of $\mathbf{e}$, and possible properties ϕ, ψ for these components, assign subjective probabilities according to the rule

$$\mathrm{prob}(\phi[\mathbf{e}(i)] \wedge \psi[\mathbf{e}(j)]) = \mathrm{prob}(\phi[\mathbf{e}(i)]) \cdot \mathrm{prob}(\psi[\mathbf{e}(j)]). \tag{5.6}$$

Subjective probabilities that postulate dependencies between $\mathbf{e}(i)$ and $\mathbf{e}(j)$ by violating (5.6) then always have to be inferred from evidence.

The second question posed above is a bit more difficult to answer. While our epistemic analysis in section 3.3 has been very general – not making restrictive assumptions about the domain of possible events of which the observed event is an instance, or the number of properties by which these events can be distinguished, and for which statistical information and prior degrees of belief are available – the statistical model of section 3.4 only has been developed for the simplified situation where events are only distinguished with respect to finitely many properties.

It is not immediately clear that this model is adequate for situations described by L^{β}-sentences, because in $\mathrm{L}^{\beta}_{\mathrm{S},\mathbf{e}}$ (for sufficiently rich S) infinitely many non-equivalent properties are definable for $\mathbf{e}$.

We might attempt to generalize the statistical model by using random variables whose range is the (infinite) set of possible events. This can be done when we presume that this set of possible events is equipped with a suitable σ-algebra, and the statistical measure μ is a

σ-measure on that σ-algebra. More sophisticated Sanov-theorems than 3.4.9 can then be used to obtain results analogous to 3.4.6, using the general definition (3.7) of cross-entropy.

However, when the set of possible events M only is equipped with an algebra and a finitely additive statistical measure μ, as is the case in a belief S-structure, we no longer even have the concept of a random variable with distribution μ in M at our disposal.

Rather than generalizing the statistical model of section 3.4, we shall therefore argue that it is in fact sufficient for the interpretation of $\phi \in \mathrm{L}^\beta$.

For this purpose, consider again the situation described in postulate 3 for the special case that the set Φ of prior degrees of belief is encoded in a sentence $\phi \in \mathrm{L}^\beta_{\mathrm{S},\mathbf{e}}$. We are then dealing with a situation in which prior degrees of belief only concern finitely many of the definable properties of $\mathbf{e}$: suppose that ϕ contains n subjective probability terms $\mathrm{prob}(\psi_i[\mathbf{e}])$ $(i = 1, \ldots, n)$. The relative frequencies of properties $\zeta(\boldsymbol{v}) \in \mathrm{L}^\beta_{\mathrm{S},\mathbf{e}}$ then are (approximately) within the bounds prescribed by ϕ, iff the relative frequencies of the properties $\psi_1(\boldsymbol{v}), \ldots, \psi_n(\boldsymbol{v})$ are (approximately) within the prescribed bounds. In other words, only the frequencies of finitely many attributes of the elements in the random sample must be checked in order to verify that the sample is consistent with our prior beliefs.

Observe that this is another consequence of not permitting free variables inside the prob()-operator. With free variables, by a single sentence, constraints for infinitely many properties can be formulated. In order to verify that a random sample satisfies the constraint

$$\forall w \ \mathrm{prob}(\mathsf{R}\, we) \geq 0.2 \quad (\in \mathscr{L}_3^=, \ \notin \mathrm{L}^\beta),$$

for instance, it must be checked for each of the (usually infinitely many) possible events e' that the relative frequency of elements x in the random sample that satisfy $\mathsf{R}\, e'x$ exceeds 0.2.

Now suppose we want to determine the subjective probability for some other property $\chi(\boldsymbol{v}) \in \mathrm{L}^\beta$. Ultimately, we are then only interested in the expected relative frequency of χ in our random samples. Thus, for the assessment of $\mathrm{prob}(\chi[\mathbf{e}])$, elements of the sample need only be distinguished by the finitely many properties $\psi_1, \ldots, \psi_n$ (for identifying the sample as accordant with our prior beliefs) and χ (for defining the new belief). Other properties ζ that elements in the sample may or may not possess, for the task at hand, are regarded as irrelevant.

For an illustration, once again, consider the film-example 3.2.2: suppose there is also a predicate CS in our language designating a film's property of having been a commercial success. We may also assume that ϕ contains the statistical information

$$[\mathsf{HE}v \mid \mathsf{CS}v]_v = 0.85.$$

We have obtained no evidence bearing on the film's likelihood of having been a commercial success, i.e. the evidence of the scene we saw on television is exhausted with regard to the property CS by the trivial bound $\mathrm{prob}(\mathsf{CS}\,\mathrm{f}) \in [0, 1]$. In spite of the availability of statistical data relating the property HE to CS, it will then be unnecessary to distinguish elements in the random sample with respect to their having been a commercial success. This property, for the inference about $\mathrm{prob}(\mathsf{HE}\,\mathrm{f})$ is irrelevant.

Note that this is a notion of irrelevance complementary to the one introduced in section 3.3. There (p. 73) a property that has been observed in an event was called irrelevant, when there is no statistical information available relating to this property. Here we argued that a property for which there is statistical information is irrelevant, when no observation inducing some degree of belief for this property has been made.

By ignoring all L^σ-definable properties except $\psi_1, \ldots, \psi_n, \chi$, we can adopt the statistical model of section 3.4 to derive $\mathrm{prob}(\chi[\mathbf{e}])$ according to postulate 3.

From modeling the random sample by a sequence of random variables with values in

$$\Gamma := \{\bigwedge_{i=1}^{n} \tilde{\psi}_i \wedge \tilde{\chi} \mid \tilde{\psi}_i \in \{\psi_i, \neg\psi_i\}, \tilde{\chi} \in \{\chi, \neg\chi\}\}$$

distributed according to the restriction to Γ of the statistical measure μ, by theorem 3.4.6 then we obtain as a condition for the subjective probability $\nu(\chi[\mathbf{e}])$ for $\mathbf{e}$ to have property χ:

$$\nu(\chi[\mathbf{e}]) \in \Pi_{\Delta(\phi)\restriction\Gamma}(\mu\restriction\Gamma)(\chi) \tag{5.7}$$

where $\Delta(\phi)\restriction\Gamma$ is $\{\nu\restriction\Gamma \mid \nu \in \Delta(\phi)\}$.

From (5.7) a partial description of a default model for ϕ is obtained. Transferring (5.7) from the abstract syntactical level to a statement about probability distributions on the domain of possible events in the framework of a belief structure $(\mathfrak{M},\nu_{\mathbf{e}})$, we obtain a condition for a preferred belief measure $\nu_{\mathbf{e}}$:

$$\nu_{\mathbf{e}}(\mathfrak{M}(\chi(\boldsymbol{v}))) \in \Pi_{\Delta_{\mathrm{F}}(\phi,\mathfrak{M})\restriction\mathfrak{A}^\chi}(\mu_{\mathbf{e}}\restriction\mathfrak{A}^\chi)(\mathfrak{M}(\chi(\boldsymbol{v}))) \tag{5.8}$$

where $\mathfrak{A}^\chi$ is the finite subalgebra of $\mathfrak{A}_{\mathbf{e}}$ generated by the interpretations $\{\mathfrak{M}(\alpha(\boldsymbol{v})) \mid \alpha \in \Gamma\}$.

(5.8) states a condition for a preferred belief measure $\nu_{\mathbf{e}}$ that should be true for any $\chi(\boldsymbol{v}) \in \mathrm{L}^\beta_{\mathrm{S},\mathbf{e}}$, and the corresponding algebra $\mathfrak{A}^\chi$ defined by $\psi_1, \ldots, \psi_n$ and χ.

The question therefore is: does there exist in $\Delta_{\mathrm{F}}(\phi,\mathfrak{M})$ a measure $\nu_{\mathbf{e}}$, such that (5.8) holds for all $\chi(\boldsymbol{v})$? If so, how can it be defined explicitly?

By theorem 5.4.2 below it will be shown that such $\nu_{\mathbf{e}}$ are obtainable by the canonical construction given in the following definition.

Definition 5.4.1 Let $\mathfrak{A}$ be an algebra, $\mu \in \Delta_{\mathrm{F}}\mathfrak{A}$. Let $\mathfrak{A}' \subseteq \mathfrak{A}$ a finite subalgebra with atoms $\{A_1, \ldots, A_p\}$ and $\nu \in \Delta_{\mathrm{F}}\mathfrak{A}'$ such that $\nu \ll \mu\restriction\mathfrak{A}'$. The extension ν^* of ν to $\mathfrak{A}$ defined by

$$\nu^*(A) := \sum_{\substack{i=1 \\ \mu(A_i) > 0}}^{p} \nu(A_i)\mu(A \mid A_i) \qquad (A \in \mathfrak{A})$$

is called the *Jeffrey-extension* of ν to $\mathfrak{A}$ by μ, denoted by $\mathcal{J}(\nu, \mu, \mathfrak{A})$.

By the next theorem, Jeffrey-extensions are adequate generalizations of CE-projections $\Pi_J(\mu)$ for measures on infinite algebras, provided that J is sufficiently simple.

Theorem 5.4.2 Let $\mathfrak{A}$ be an algebra, $\mu \in \Delta_{\mathrm{F}}\mathfrak{A}$. Let $J \subseteq \Delta_{\mathrm{F}}\mathfrak{A}$ be defined by constraints on a finite subalgebra $\mathfrak{A}' \subseteq \mathfrak{A}$, i.e.

$$\forall \nu \in \Delta_{\mathrm{F}}\mathfrak{A}: \quad \nu \in J \;\Leftrightarrow\; \nu\restriction\mathfrak{A}' \in J\restriction\mathfrak{A}'. \tag{5.9}$$

Then for all finite $\mathfrak{A}'' \supseteq \mathfrak{A}'$:

$$\Pi_{J\upharpoonright\mathfrak{A}''}(\mu\upharpoonright\mathfrak{A}'') = \{\nu\upharpoonright\mathfrak{A}'' \mid \nu = \mathcal{J}(\nu', \mu, \mathfrak{A}),\ \nu' \in \Pi_{J\upharpoonright\mathfrak{A}'}(\mu\upharpoonright\mathfrak{A}')\}. \tag{5.10}$$

Conversely, for $\nu \in \Delta_{\mathrm{F}}\mathfrak{A}$, if

$$\nu\upharpoonright\mathfrak{A}'' \in \Pi_{J\upharpoonright\mathfrak{A}''}(\mu\upharpoonright\mathfrak{A}'') \tag{5.11}$$

for all finite $\mathfrak{A}'' \supseteq \mathfrak{A}'$, then $\nu = \mathcal{J}(\nu', \mu, \mathfrak{A})$ for some $\nu' \in \Pi_{J\upharpoonright\mathfrak{A}'}(\mu\upharpoonright\mathfrak{A}')$.

Proof: Let $\{A_1, \ldots, A_p\}$ be the set of atoms of $\mathfrak{A}'$. Let $\nu'' \in \Delta_{\mathrm{F}}\mathfrak{A}''$, $\nu'' \ll \mu\upharpoonright\mathfrak{A}''$. By lemma 4.0.16 then

$$CE^{\mathrm{F}}(\nu'', \mu\upharpoonright\mathfrak{A}'') \geq CE^{\mathrm{F}}(\nu''\upharpoonright\mathfrak{A}', \mu\upharpoonright\mathfrak{A}')$$

with equality iff

$$(\nu'')^h = (\mu\upharpoonright\mathfrak{A}'')^h \qquad (h = 1, \ldots, p) \tag{5.12}$$

where $(\cdot)^h$ is the conditional distribution on A_h. Equivalent to (5.12) is

$$\nu'' = \mathcal{J}(\nu''\upharpoonright\mathfrak{A}', \mu\upharpoonright\mathfrak{A}'', \mathfrak{A}'').$$

Since by (5.9)

$$\mathcal{J}(\nu', \mu\upharpoonright\mathfrak{A}'', \mathfrak{A}'') \in J\upharpoonright\mathfrak{A}''$$

for all $\nu' \in J\upharpoonright\mathfrak{A}'$, we obtain

$$\Pi_{J\upharpoonright\mathfrak{A}''}(\mu\upharpoonright\mathfrak{A}'') = \{\mathcal{J}(\nu', \mu\upharpoonright\mathfrak{A}'', \mathfrak{A}'') \mid \nu' \in \Pi_{J\upharpoonright\mathfrak{A}'}(\mu\upharpoonright\mathfrak{A}')\}. \tag{5.13}$$

With

$$\mathcal{J}(\nu', \mu\upharpoonright\mathfrak{A}'', \mathfrak{A}'') = \mathcal{J}(\nu', \mu, \mathfrak{A})\upharpoonright\mathfrak{A}''$$

this proves (5.10).

Now assume that (5.11) holds for $\nu \in \Delta_{\mathrm{F}}\mathfrak{A}$ and all finite $\mathfrak{A}'' \supseteq \mathfrak{A}'$. Specifically, then

$$\nu\upharpoonright\mathfrak{A}' := \nu' \in \Pi_{J\upharpoonright\mathfrak{A}'}(\mu\upharpoonright\mathfrak{A}').$$

Let $A \in \mathfrak{A}$ be arbitrary and $\mathfrak{A}''$ the algebra generated by $\mathfrak{A}'$ and A. According to (5.11) and (5.13)

$$\nu(A) = \mathcal{J}(\nu', \mu\upharpoonright\mathfrak{A}'', \mathfrak{A}''),$$

and hence $\nu = \mathcal{J}(\nu', \mu, \mathfrak{A})$. $\square$

The cross-entropy minimization properties of Jeffrey-extensions $\mathcal{J}(\nu, \mu, \mathfrak{A})$ for $\nu \in \Pi_{J\upharpoonright\mathfrak{A}'}(\mu\upharpoonright\mathfrak{A}')$ suggest to extend our previous notation, and to write

$$\Pi_J(\mu, \mathfrak{A}') := \{\mathcal{J}(\nu, \mu, \mathfrak{A}) \mid \nu \in \Pi_{J\upharpoonright\mathfrak{A}'}(\mu\upharpoonright\mathfrak{A}')\}. \tag{5.14}$$

However, the analogy between the set designated by (5.14), and our older concept of the CE-projection of μ into J as introduced in 3.4.1, appears to be somewhat unsatisfactory, because of the former's dependence on the algebra $\mathfrak{A}'$ used to represent J, which certainly will not be the unique algebra with the property (5.9): for every $\mathfrak{A}'' \supseteq \mathfrak{A}'$, for instance, (5.9) also holds with $\mathfrak{A}''$ in place of $\mathfrak{A}'$.

To see that the set $\Pi_J(\mu, \mathfrak{A}')$ really extends the concept of the CE-projection of μ into J, we will therefore have to show that it is in fact independent of the particular choice of $\mathfrak{A}'$.

When $\mathfrak{A}'' \supseteq \mathfrak{A}'$, it is not difficult to see that $\Pi_J(\mu, \mathfrak{A}') = \Pi_J(\mu, \mathfrak{A}'')$, because

$$\begin{aligned}
\Pi_J(\mu, \mathfrak{A}'') &= \{\mathcal{J}(\nu, \mu, \mathfrak{A}) \mid \nu \in \Pi_{J\restriction\mathfrak{A}''}(\mu\restriction\mathfrak{A}'')\} \\
&= \{\mathcal{J}(\mathcal{J}(\nu', \mu\restriction\mathfrak{A}'', \mathfrak{A}''), \mu, \mathfrak{A}) \mid \nu' \in \Pi_{J\restriction\mathfrak{A}'}(\mu\restriction\mathfrak{A}')\} \qquad \text{by (5.13)} \\
&= \{\mathcal{J}(\nu', \mu, \mathfrak{A}) \mid \nu' \in \Pi_{J\restriction\mathfrak{A}'}(\mu\restriction\mathfrak{A}')\} \\
&= \Pi_J(\mu, \mathfrak{A}').
\end{aligned}$$

What if J can be defined by constraints on $\mathfrak{A}'$ and $\mathfrak{A}''$, but neither $\mathfrak{A}' \subseteq \mathfrak{A}''$, nor $\mathfrak{A}'' \subseteq \mathfrak{A}'$? Does $\Pi_J(\mu, \mathfrak{A}') = \Pi_J(\mu, \mathfrak{A}'')$ still hold? A positive answer to this question is provided by the following lemma, which states that in this situation J also is determined by constraints on $\mathfrak{A}^\cap := \mathfrak{A}' \cap \mathfrak{A}''$. By $\mathfrak{A}^\cap \subseteq \mathfrak{A}'$, $\mathfrak{A}^\cap \subseteq \mathfrak{A}''$, and our considerations above, this implies that

$$\Pi_J(\mu, \mathfrak{A}') = \Pi_J(\mu, \mathfrak{A}^\cap) = \Pi_J(\mu, \mathfrak{A}'').$$

Lemma 5.4.3 Let $\mathfrak{A}$ be an algebra, $\mathfrak{A}', \mathfrak{A}''$ finite subalgebras of $\mathfrak{A}$, and $J \subseteq \Delta_{\mathrm{F}}\mathfrak{A}$ such that

$$\forall \nu \in \Delta_{\mathrm{F}}\mathfrak{A}: \quad \nu \in J \quad \Leftrightarrow \quad \nu\restriction\mathfrak{A}' \in J\restriction\mathfrak{A}' \quad \Leftrightarrow \quad \nu\restriction\mathfrak{A}'' \in J\restriction\mathfrak{A}''. \tag{5.15}$$

With

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

then

$$\forall \nu \in \Delta_{\mathrm{F}}\mathfrak{A}: \quad \nu \in J \quad \Leftrightarrow \quad \nu\restriction\mathfrak{A}^\cap \in J\restriction\mathfrak{A}^\cap. \tag{5.16}$$

Proof: Let $\mathfrak{A}^\cup$ be the subalgebra of $\mathfrak{A}$ generated by $\mathfrak{A}'$ and $\mathfrak{A}''$. Since (5.15) then implies that

$$\forall \nu \in \Delta_{\mathrm{F}}\mathfrak{A}: \quad \nu \in J \quad \Leftrightarrow \quad \nu\restriction\mathfrak{A}^\cup \in J\restriction\mathfrak{A}^\cup,$$

for (5.16) it suffices to show that

$$\nu\restriction\mathfrak{A}^\cup \in J\restriction\mathfrak{A}^\cup \quad \Leftrightarrow \quad \nu\restriction\mathfrak{A}^\cap \in J\restriction\mathfrak{A}^\cap. \tag{5.17}$$

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

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 $\mathfrak{A}^\cup$ then 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 $J \subseteq \{1, \ldots, q\}$, such that

$$A = \bigcup_{j \in J} A''_j.$$

A is an atom of $\mathfrak{A}^\cap$ iff the sets I, J defining A are minimal, i.e. there are no $I' \subset I$, $J' \subset J$ such that

$$\bigcup_{i \in I'} A'_i = \bigcup_{j \in J'} A''_j.$$

Figure 5.1 depicts the four algebras $\mathfrak{A}^\cap, \mathfrak{A}', \mathfrak{A}'', \mathfrak{A}^\cup$. In this picture, atoms of $\mathfrak{A}'$ are represented by fields delimited by dotted lines, atoms of $\mathfrak{A}''$ by fields delimited by dashed lines. Solid lines mark the boundaries of atoms of $\mathfrak{A}^\cap$, and each field defined by lines of any kind represents an atom of $\mathfrak{A}^\cup$.

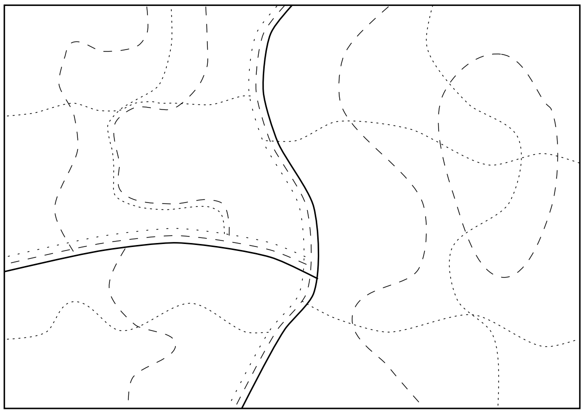

Figure 5.1: Algebras in the proof of lemma 5.4.3

Figure 5.1 also may be used to explain the basic idea for the proof of (5.17).

A measure $\nu \in \Delta_{\mathrm{F}}\mathfrak{A}^\cup$ is given by an assignment of probabilities to the atoms of $\mathfrak{A}^\cup$. Consider a measure ν' on $\mathfrak{A}^\cup$ obtained from $\nu \in \Delta_{\mathrm{F}}\mathfrak{A}^\cup$ by shifting part of the probability $\nu(B_{ij})$ of some atom B_{ij} to another atom $B_{ij'}$ which is a subset of the same $A'_i \in \mathfrak{A}'$ as B_{ij}. More precisely: let $r \leq \nu(B_{ij})$, and define

$$\nu'(B_{ij}) := \nu(B_{ij}) - r, \quad \nu'(B_{ij'}) := \nu(B_{ij'}) + r,$$

and $\nu'(B) := \nu(B)$ for all other atoms B of $\mathfrak{A}^\cup$. Then the restricted distribution of ν' on $\mathfrak{A}'$ has not changed from the one of ν, so that by (5.15) we have that

$$\nu' \in J \;\Leftrightarrow\; \nu \in J.$$

This process can be iterated: the probability weight r can be shifted from $B_{ij'}$ to another atom of $\mathfrak{A}^\cup$, which is either of the form $B_{ij''}$, i.e. belonging to the same atom of $\mathfrak{A}'$ as $B_{ij'}$, or of the form $B_{i'j'}$, i.e. belonging to the same atom of $\mathfrak{A}''$ as $B_{ij'}$. In both cases a new measure ν'' is obtained that has the same restricted distribution as ν' on either $\mathfrak{A}'$ or $\mathfrak{A}''$. Thus,

$$\nu'' \in J \;\Leftrightarrow\; \nu' \in J.$$

To prove the lemma, we will show that in this manner the weight r can in fact be shifted from B_{ij} to any atom $B_{i'j'}$ of $\mathfrak{A}^{\cup}$ belonging to the same atom of $\mathfrak{A}^{\cap}$ as B_{ij}. The result then follows easily, for assume that $\nu \in \Delta_{\mathrm{F}}\mathfrak{A}^{\cup}$ with $\nu\restriction\mathfrak{A}^{\cap} \in J\restriction\mathfrak{A}^{\cap}$, so that there exists $\tilde{\nu} \in J$ with $\tilde{\nu}\restriction\mathfrak{A}^{\cap} = \nu\restriction\mathfrak{A}^{\cap}$. Clearly, ν can be transformed into $\tilde{\nu}$ by a finite series of weight shifts inside atoms of $\mathfrak{A}^{\cap}$. Each of these shifts preserving membership in J, we obtain that $\nu \in J$, which proves (5.16).

It remains to be shown that it is indeed possible to shift probability weights among arbitrary atoms of $\mathfrak{A}^{\cup}$ inside atoms of $\mathfrak{A}^{\cap}$ while preserving membership of J.

Intuitively, this is fairly obvious by a look at figure 5.1: inside atoms of $\mathfrak{A}^{\cap}$, each pair of atoms of $\mathfrak{A}^{\cup}$ can be connected by a "path" which only crosses boundaries of either atoms of $\mathfrak{A}'$, or of $\mathfrak{A}''$, but not simultaneously boundaries of both kinds. To formalize this argument, we introduce a binary relation $\sim$ on atoms of $\mathfrak{A}^{\cup}$ by

$$B_{ij} \sim B_{i'j'} \;\text{ iff }\; i = i' \text{ or } j = j'.$$

Let $\approx$ be the transitive closure of $\sim$. Clearly, $\approx$ is an equivalence relation on the atoms of $\mathfrak{A}^{\cup}$. Let $[B_{ij}]$ denote the equivalence class of B_{ij} with respect to $\approx$. It then must be shown that $\cup[B_{ij}]$ contains the atom of $\mathfrak{A}^{\cap}$ that B_{ij} is a subset of. For this we only need to show that $\cup[B_{ij}]$ is an element of $\mathfrak{A}^{\cap}$.

For B_{ij} define

$$\begin{aligned} I([B_{ij}]) &:= \{i' \in \{1,\dots,p\} \mid \exists j' \in \{1,\dots,q\}\; B_{i'j'} \in [B_{ij}]\} \\ J([B_{ij}]) &:= \{j' \in \{1,\dots,q\} \mid \exists i' \in \{1,\dots,p\}\; B_{i'j'} \in [B_{ij}]\}. \end{aligned}$$

For all $i' \in I([B_{ij}])$ then

$$A'_{i'} \subseteq \cup[B_{ij}], \tag{5.18}$$

because $A'_{i'} = \cup_{j' \in \{1,\dots,q\}} B_{i'j'}$, $B_{i'j'} \subseteq \cup[B_{ij}]$ for some j', and $B_{i'j''} \sim B_{i'j'}$ for all other $j'' \in \{1,\dots,q\}$. On the other hand, obviously,

$$\cup\,[B_{ij}] \subseteq \cup\{A'_{i'} \mid i' \in I\}. \tag{5.19}$$

Together with the analogous statements about atoms A''_j of $\mathfrak{A}''$, (5.18) and (5.19) yield

$$\bigcup_{i' \in I([B_{ij}])} A'_{i'} \;=\; \bigcup_{j' \in J([B_{ij}])} A'_{i'} \;=\; \cup[B_{ij}] \in \mathfrak{A}^{\cap}.$$

$\square$

Lemma 5.4.3 justifies the following definition.

Definition 5.4.4 Let $\mathfrak{A}$, μ, J, and $\mathfrak{A}'$ as in theorem 5.4.2. We write

$$\Pi_J(\mu) := \{\mathcal{J}(\nu, \mu, \mathfrak{A}) \mid \nu \in \Pi_{J\upharpoonright\mathfrak{A}'}(\mu\upharpoonright\mathfrak{A}')\}.$$

With theorem 5.4.2 and definition 5.4.4 it becomes clear how to define a set of preferred belief measures satisfying (5.8) for every set $\mathfrak{M}(\chi(\boldsymbol{v}))$.

Letting $\mathfrak{A} = \mathfrak{A}_{\mathbf{e}}$, $\mu = \mu_{\mathbf{e}}$, and $J = \Delta_{\mathrm{F}}(\phi, \mathfrak{M})$, theorem 5.4.2 is applicable, because $\Delta_{\mathrm{F}}(\phi, \mathfrak{M})$ is defined by constraints on the finite subalgebra of $\mathfrak{A}_{\mathbf{e}}$ generated by the extensions of $\psi_1(\boldsymbol{v}), \ldots, \psi_n(\boldsymbol{v})$. By theorem 5.4.2, (5.8) is satisfied for every $\nu_{\mathbf{e}} \in \Pi_{\Delta_{\mathrm{F}}(\phi,\mathfrak{M})}(\mu_{\mathbf{e}})$. Conversely, every measure $\nu_{\mathbf{e}}$ that satisfies (5.8) for all $\chi(\boldsymbol{v})$ belongs to $\Pi_{\Delta_{\mathrm{F}}(\phi,\mathfrak{M})}(\mu_{\mathbf{e}})$.

Definition 5.4.5 Let $\phi \in \mathrm{L}^{\beta}_{\mathrm{S},\mathbf{e}}$, $(\mathfrak{M}, \nu_{\mathbf{e}})$ a feasible model of ϕ. $(\mathfrak{M}, \nu_{\mathbf{e}})$ is called a *default model* of ϕ, written $(\mathfrak{M}, \nu_{\mathbf{e}}) \mathrel{|\!\!\approx} \phi$, iff $\nu_{\mathbf{e}} \in \Pi_{\Delta_{\mathrm{F}}(\phi,\mathfrak{M})}(\mu_{\mathbf{e}})$.

Observe that this definition can not be extended easily to infinite theories $\Phi \subseteq \mathrm{L}^{\beta}_{\mathrm{S},\mathbf{e}}$, because such Φ define sets $\Delta_{\mathrm{F}}(\Phi, \mathfrak{M})$ of feasible belief measures not definable by constraints on a finite subalgebra of $\mathfrak{A}_{\mathbf{e}}$, so that we do not have a useful definition of a set $\Pi_{\Delta_{\mathrm{F}}(\Phi,\mathfrak{M})}(\mu_{\mathbf{e}})$ of preferred belief measures.

Example 5.4.6 Suppose that $\phi \in \mathrm{L}^{\sigma}_{\mathrm{S}}$ and $\mathfrak{M} \models_{\beta} \phi$ for a statistical base structure $\mathfrak{M}$. Then $(\mathfrak{M}, \nu_{\mathbf{e}})$ is a feasible model of ϕ for all $\nu_{\mathbf{e}} \in \Delta_{\mathrm{F}}\mathfrak{A}_{\mathbf{e}}$:

$$\Delta_{\mathrm{F}}(\phi, \mathfrak{M}) = \Delta_{\mathrm{F}}\mathfrak{A}_{\mathbf{e}}.$$

Therefore $\Pi_{\Delta_{\mathrm{F}}(\phi,\mathfrak{M})}(\mu_{\mathbf{e}}) = \{\mu_{\mathbf{e}}\}$, so that

$$(\mathfrak{M}, \nu_{\mathbf{e}}) \mathrel{|\!\!\approx} \phi \;\text{ iff }\; \nu_{\mathbf{e}} = \mu_{\mathbf{e}}. \tag{5.20}$$

Definition 5.4.7 Let $\phi, \psi \in \mathrm{L}^{\beta}$. ψ is *default entailed* by ϕ (written $\phi \mathrel{|\!\!\approx} \psi$) iff $(\mathfrak{M}, \nu_{\mathbf{e}}) \models_{\beta} \psi$ for all default models $(\mathfrak{M},\nu_{\mathbf{e}})$ of ϕ. The relativized operator $\mathrel{|\!\!\approx}^{\mathbf{R}}$ is used for default entailment with respect to real-valued probabilities, i.e. $\phi \mathrel{|\!\!\approx}^{\mathbf{R}} \psi$ iff $(\mathfrak{M}, \nu_{\mathbf{e}}) \models_{\beta} \psi$ for every real-valued default model $(\mathfrak{M},\nu_{\mathbf{e}})$ of ϕ.

Definition 5.4.8 We write $\mathscr{L}^{\beta}_{\mathrm{def}}$ for the logic defined by the language L^{β} and the entailment relation $\mathrel{|\!\!\approx}$.

When analogous statements can be made for both $\mathrel{|\!\!\approx}$ and $\mathrel{|\!\!\approx}^{\mathbf{R}}$, we use the denotation $\mathrel{|\!\!\approx}^{(\mathbf{R})}$, which means that the statement in which $\mathrel{|\!\!\approx}^{(\mathbf{R})}$ occurs is true when $\mathrel{|\!\!\approx}^{(\mathbf{R})}$ is either replaced by $\mathrel{|\!\!\approx}$, or by $\mathrel{|\!\!\approx}^{\mathbf{R}}$ throughout.

Example 5.4.9 Let $\phi, \chi(\boldsymbol{v}) \in \mathrm{L}^{\sigma}$, $(\mathfrak{M}, \nu_{\mathbf{e}}) \mathrel{|\!\!\approx} \phi$. By example 5.4.6 then $\nu_{\mathbf{e}} = \mu_{\mathbf{e}}$, so that $\mathfrak{M} \models_{\beta} \mathrm{prob}(\chi[\mathbf{e}]) = [\chi(\boldsymbol{v})]_{\boldsymbol{v}}$. Hence

$$\phi \mathrel{|\!\!\approx}^{(\mathbf{R})} \mathrm{prob}(\chi[\mathbf{e}]) = [\chi(\boldsymbol{v})]_{\boldsymbol{v}}.$$

Clearly, $\approx^{(\mathbf{R})}$ is a nonmonotonic inference relation: adding new constraints for prior degrees of belief usually invalidates former default inferences. This, however, is the only way in which $\approx^{(\mathbf{R})}$ is nonmonotonic. As the following lemma points out, $\approx^{(\mathbf{R})}$ behaves monotonically for deterministic formulas from L^{σ}.

Lemma 5.4.10 Let $\phi, \psi \in \mathrm{L}^{\beta}$ with $\phi \approx^{(\mathbf{R})} \psi$. Let $\chi \in \mathrm{L}^{\sigma}$. Then $\phi \wedge \chi \approx^{(\mathbf{R})} \psi$.

Proof: Let $(\mathfrak{M}, \nu_{\mathbf{e}})$ be a default model of $\phi \wedge \chi$. Since for all $\nu'_{\mathbf{e}} \in \Delta_{\mathrm{F}}\mathfrak{A}_{\mathbf{e}}$, $(\mathfrak{M}, \nu'_{\mathbf{e}})$ is a feasible model of $\phi \wedge \chi$ iff $(\mathfrak{M}, \nu'_{\mathbf{e}})$ is a feasible model of ϕ (cf. example 5.4.6), we have that

$$\Delta_{\mathrm{F}}(\phi \wedge \chi, \mathfrak{M}) = \Delta_{\mathrm{F}}(\phi, \mathfrak{M}).$$

Hence

$$\nu_{\mathbf{e}} \in \Pi_{\Delta_{\mathrm{F}}(\phi,\mathfrak{M})}(\mu_{\mathbf{e}}) = \Pi_{\Delta_{\mathrm{F}}(\phi\wedge\chi,\mathfrak{M})}(\mu_{|\mathbf{e}|}),$$

and $(\mathfrak{M}, \nu_{\mathbf{e}}) \approx^{(\mathbf{R})} \phi$. Thus, $(\mathfrak{M}, \nu_{\mathbf{e}}) \models_{\beta} \psi$. □

For $\approx^{\mathbf{R}}$ it is once more easy to show incompleteness by reducing $\models^{\mathbf{R}}_{\beta}$ to $\approx^{\mathbf{R}}$.

Lemma 5.4.11 Let $\phi, \psi \in \mathrm{L}^{\sigma}$. Then

$$\phi \models^{\mathbf{R}}_{\beta} \psi \;\Leftrightarrow\; \phi \approx^{\mathbf{R}} \psi.$$

Proof: The left to right direction is immediate from the definitions. For the right to left direction let $(\mathfrak{M},\nu_{\mathbf{e}})$ be a real-valued feasible model of ϕ. Then $(\mathfrak{M},\mu_{\mathbf{e}})$ is a real-valued default model of ϕ (cf. example 5.4.6), and therefore $(\mathfrak{M}, \mu_{\mathbf{e}}) \models_{\beta} \psi$. Since $\psi \in \mathrm{L}^{\sigma}$, then also $(\mathfrak{M}, \nu_{\mathbf{e}}) \models_{\beta} \psi$. □

It is perhaps worthwhile to recapitulate the various assumptions that had to be made in order to arrive at the default semantics of definition 5.4.5. These assumptions should always be borne in mind, when default inferences are computed from a formula $\phi \in \mathrm{L}^{\beta}$.

First, there is the fundamental assumption of postulate 2: the uncertain events are taken to have been produced by a random mechanism equivalent to the statistical sampling rule. As we have seen, this entails the assumption of two or more different uncertain events to be independent realizations of equivalent random mechanisms. Furthermore, there is the assumption of postulate 3: the stated prior beliefs have to be complete in the sense that they exhaust the evidence. This is a typical assumption that always has to be made when a nonmonotonic inference rule is applied to a knowledge base representing the knowledge of an intelligent agent: such an inference can necessarily only be argued to correspond to the agent's own reasoning, when it is taken for granted that the agent does not possess any additional knowledge which might invalidate the default inference.

The remainder of this section is dedicated to a partial exploration of the properties of default entailment $\approx^{(\mathbf{R})}$ by means of several examples.

In the first example we pick up the discussion of section 3.4.3 about how to deal with degrees of belief that define non-closed sets of belief measures $\Delta(\Phi)$. By definition 5.4.5 the conservative approach is taken towards interpreting sentences ϕ that give rise to non-closed

$\Delta_{\mathrm{F}}(\phi, \mathfrak{M})$: a default belief measure $\nu_{\mathbf{e}}$ must belong to $\Delta_{\mathrm{F}}(\phi, \mathfrak{M})$, not merely to $cl\, \Delta_{\mathrm{F}}(\phi, \mathfrak{M})$. The more adventurous approach of letting $\nu_{\mathbf{e}} \in \Pi_{cl\, \Delta_{\mathrm{F}}(\phi, \boldsymbol{M})}(\mu_{\mathbf{e}})$ obviously would have led to a logic with rather bizarre properties, even defying what Kraus et al. [1990] have named the reflexivity property, i.e. $\phi \mid\!\approx \phi$ for all ϕ.

Example 5.4.12 Let $\mathsf{R} \in \mathrm{S}$ be a unary predicate symbol, e an event symbol. Consider

$$\phi :\equiv [\mathsf{R}\, v]_v = 0.1 \wedge \mathrm{prob}(\mathsf{R}\, \mathrm{e}) > 0.8. \tag{5.21}$$

Let $\mathfrak{M} = (M, I, \mathfrak{F}, (\mathfrak{A}_n, \mu_n)_n)$ be a statistical base S-structure. If $\mu_1(I(\mathsf{R})) \neq 0.1$, then $\Delta_{\mathrm{F}}(\phi, \mathfrak{M}) = \emptyset$. Otherwise, $\Delta_{\mathrm{F}}(\phi, \mathfrak{M})$ is defined by constraints on the finite subalgebra $\mathfrak{A}' = \{\emptyset, I(\mathsf{R}), I(\mathsf{R})^c, M\}$ of $\mathfrak{A}_1$ by $\Delta_{\mathrm{F}}(\phi, \mathfrak{M}) \restriction \mathfrak{A}' = \{(\nu_1, \nu_2) \in \Delta^2_{\mathrm{F}} \mid \nu_1 > 0.8\}$. Since this is an open set that does not contain $\mu_{\mathrm{e}} \restriction \mathfrak{A}'$, we have that $\Pi_{\Delta_{\mathrm{F}}(\phi, \mathfrak{M})}(\mu_{\mathrm{e}}) = \emptyset$ (cf. example 3.4.3). Consequently, ϕ does not have a default model, and $\phi \mid\!\approx^{(\mathbf{R})} \kappa$ for every $\kappa \in \mathrm{L}^{\beta}$.

Compare this to

$$\phi' :\equiv [\mathsf{R}\, v]_v > 0.9 \wedge \mathrm{prob}(\mathsf{R}\, \mathrm{e}) > 0.8.$$

Here $\Delta_{\mathrm{F}}(\phi, \mathfrak{M}) = \emptyset$ if $\mu_1(I(\mathsf{R})) \leq 0.9$, and $\mu_{\mathrm{e}} \in \Delta_{\mathrm{F}}(\phi, \mathfrak{M})$ else, so that $\mu_{\mathrm{e}} = \pi_{\Delta_{\mathrm{F}}(\phi, \mathfrak{M})}(\mu_{\mathrm{e}})$. Hence, default models of ϕ' are just the feasible models of ϕ' with $\nu_{\mathrm{e}} = \mu_{\mathrm{e}}$, and

$$\phi' \mid\!\approx^{(\mathbf{R})} \mathrm{prob}(\mathsf{R}\, \mathrm{e}) > 0.9. \tag{5.22}$$

Finally consider

$$\begin{array}{rcl} \phi^2 & :\equiv & [\mathsf{R}\, v]_v = 0.5 \wedge \mathrm{prob}(\mathsf{R}\, \mathrm{e}) < 0.2 \vee \mathrm{prob}(\mathsf{R}\, \mathrm{e}) \geq 0.8 \\ \phi^3 & :\equiv & [\mathsf{R}\, v]_v = 0.5 \wedge \mathrm{prob}(\mathsf{R}\, \mathrm{e}) \leq 0.2 \vee \mathrm{prob}(\mathsf{R}\, \mathrm{e}) \geq 0.8 \\ \phi^4 & :\equiv & [\mathsf{R}\, v]_v = 0.5 \wedge \mathrm{prob}(\mathsf{R}\, \mathrm{e}) \leq 0.19 \vee \mathrm{prob}(\mathsf{R}\, \mathrm{e}) \geq 0.8 \end{array}$$

Here, for $\mathfrak{M}$ with $\mu_1(I(\mathsf{R})) = 0.5$,

$$\Pi_{\Delta_{\mathrm{F}}(\phi^2, \mathfrak{M})}(\mu_{\mathrm{e}}) = \Pi_{\Delta_{\mathrm{F}}(\phi^4, \mathfrak{M})}(\mu_{\mathrm{e}}) = \{(0.8, 0.2)\} \qquad \Pi_{\Delta_{\mathrm{F}}(\phi^3, \mathfrak{M})}(\mu_{\mathrm{e}}) = \{(0.8, 0.2), (0.2, 0.8)\},$$

so that

$$\begin{gathered} \phi^2 \mid\!\approx^{(\mathbf{R})} \mathrm{prob}(\mathsf{R}\, \mathrm{e}) = 0.8, \quad \phi^4 \mid\!\approx^{(\mathbf{R})} \mathrm{prob}(\mathsf{R}\, \mathrm{e}) = 0.8, \\ \phi^3 \mid\!\approx^{(\mathbf{R})} \mathrm{prob}(\mathsf{R}\, \mathrm{e}) = 0.8 \vee \mathrm{prob}(\mathsf{R}\, \mathrm{e}) = 0.2. \end{gathered} \tag{5.23}$$

This example demonstrates that one must be very careful when default inferences are drawn from sentences ϕ for which $\Delta_{\mathrm{F}}(\phi, \mathfrak{M})$ may be non-closed. (5.22), on the other hand, shows that sentences ϕ with $\Delta(\phi, \mathfrak{M})$ an open set, do not necessarily give rise only to inconsistent or useless default inferences.

The following three examples contain a treatment of our two ubiquitous examples 3.2.1 and 3.2.2 within the default semantics for L^{β}.

Example 5.4.13 The situation described in example 3.2.1 may be encoded by the conjunction ϕ of the L^{β}-sentences

$$\forall v (\mathsf{R}_1 v \dot{\vee} \ldots \dot{\vee} \mathsf{R}_6) \tag{5.24}$$

$$\bigwedge_{i=1}^{6} [\mathsf{R}_i v]_v = 1/6 \tag{5.25}$$

$$\mathrm{prob}(\mathsf{R}_1\, \mathrm{t} \vee \mathsf{R}_2\, \mathrm{t} \vee \mathsf{R}_3\, \mathrm{t}) = 0.3. \tag{5.26}$$

Suppose that $(\mathfrak{M}, \nu_{\mathrm{t}})$ is a feasible model for ϕ. The interpretations $I(\mathsf{R}_i)$ $(i = 1, \ldots, 6)$ then form a partition of M with $\mu_1(I(\mathsf{R}_i)) = 1/6$. The set $\Delta(\phi, \mathfrak{M})$ is determined by constraints on the finite subalgebra $\mathfrak{A}'$ of $\mathfrak{A}_{\mathrm{t}}$ given by the six atoms $I(\mathsf{R}_i)$. Writing ν_i for $\nu(I(\mathsf{R}_i))$ $(\nu \in \Delta\mathfrak{A}_{\mathrm{t}})$, we obtain

$$\Delta(\phi, \mathfrak{M}) \restriction \mathfrak{A}' = \{(\nu_1, \ldots, \nu_6) \mid \nu_1 + \nu_2 + \nu_3 = 0.3\}.$$

Then

$$\pi_{\Delta(\phi,\mathfrak{M})}(\mu_{\mathrm{t}}) \restriction \mathfrak{A}' = (0.1, 0.1, 0.1, 0.7/3, 0.7/3, 0.7/3),$$

and $(\mathfrak{M}, \nu_{\mathrm{t}})$ is a default model of ϕ only if $\nu_{\mathrm{t}} \restriction \mathfrak{A}' = (0.1, \ldots, 0.7/3)$. Hence

$$\begin{array}{ll} \phi \mid\!\approx^{(\mathbf{R})} \mathrm{prob}(\mathsf{R}_i\,\mathrm{t}) = 0.1 & (i = 1, 2, 3) \\ \phi \mid\!\approx^{(\mathbf{R})} \mathrm{prob}(\mathsf{R}_i\,\mathrm{t}) = 0.7/3 & (i = 4, 5, 6). \end{array} \tag{5.27}$$

Example 5.4.14 To formalize the modified die-example incorporating the possibility that a loaded die has been cast (cf. p. 65), let

$$\sigma_1 \quad :\equiv \quad \bigwedge_{i=1}^{6} [\mathsf{R}_i v]_v = 1/6 \tag{5.28}$$

$$\sigma_2 \quad :\equiv \quad [\mathsf{R}_1 v]_v = 0.012 \wedge \bigwedge_{i=2}^{5} [\mathsf{R}_i v]_v = 0.022 \wedge [\mathsf{R}_6 v]_v = 0.9. \tag{5.29}$$

Then the situation described in the example may be represented by (5.24) and the following three formulas.

$$\sigma_1 \vee \sigma_2 \tag{5.30}$$

$$\sigma_1 \to \mathrm{prob}(\mathsf{R}_1\,\mathrm{t} \vee \mathsf{R}_2\,\mathrm{t} \vee \mathsf{R}_3\,\mathrm{t}) = 0.3 \tag{5.31}$$

$$\sigma_2 \to \mathrm{prob}(\mathsf{R}_1\,\mathrm{t} \vee \mathsf{R}_2\,\mathrm{t} \vee \mathsf{R}_3\,\mathrm{t}) = 0.01. \tag{5.32}$$

Let ϕ be the conjunction of (5.24) and (5.30)-(5.32). For statistical base S-structures $\mathfrak{M}$ with μ_1 according to the constraints in σ_1 we have

$$\pi_{\Delta_{\mathrm{F}}(\phi,\mathfrak{M})}(\mu_{\mathrm{t}})(I(\mathsf{R}_1)) = 0.1$$

(as before, by Jeffrey's rule), while

$$\pi_{\Delta_{\mathrm{F}}(\phi,\mathfrak{M})}(\mu_{\mathrm{t}})(I(\mathsf{R}_1)) = 0.01 \frac{0.012}{0.012 + 2 \cdot 0.022} \approx 0.0021$$

when μ_{t} is according to σ_2. Hence

$$\phi \mid\!\approx^{(\mathbf{R})} (\sigma_1 \to \mathrm{prob}(\mathsf{R}_1\mathrm{t}) = 0.1) \wedge (\sigma_2 \to \mathrm{prob}(\mathsf{R}_1\mathrm{t}) \in [0.00205, 0.00215]). \tag{5.33}$$

It has been our intention in the development of the default semantics for L^{β}, that default inferences are restricted to statements of subjective beliefs, while objective statements only should be inferred when logically implied. Ideally, this would mean that for $\phi \in \mathrm{L}^{\beta}$ and objective sentences $\psi \in \mathrm{L}^{\sigma}$

$$\phi \mid\!\approx \psi \quad \text{iff} \quad \phi \models_{\beta} \psi. \tag{5.34}$$

A modification of example 5.4.14 shows that this, in general, can not be guaranteed.

Example 5.4.15 Let σ_1, σ_2 as in example 5.4.14. Modify (5.31) by replacing equality with inequality:

$$\sigma_1 \rightarrow \mathrm{prob}(\mathsf{R}_1\,\mathrm{t} \vee \mathsf{R}_2\,\mathrm{t} \vee \mathsf{R}_3\,\mathrm{t}) < 0.3. \tag{5.35}$$

Let ϕ be the conjunction of (5.24), (5.30), (5.35), and (5.32). For a statistical base structure $\mathfrak{M}$ with μ_1 according to σ_1 we now have $\Pi_{\Delta_{\mathrm{F}}(\phi,\mathfrak{M})}(\mu_{\mathrm{t}}) = \emptyset$, so that in every default model of ϕ the statistical measure μ_1 is described by σ_2. Hence

$$\phi \mid\!\approx \sigma_2, \quad \text{but} \quad \phi \not\models_\beta \sigma_2.$$

The failure of (5.34) in this example is caused by the fact that for ϕ there exist feasible models with a statistical base structure that is not the statistical base structure of any default model. For any given ϕ, for which it is ensured that for every statistical base structure $\mathfrak{M}$ we have that

$$\Delta_{\mathrm{F}}(\phi,\mathfrak{M}) \neq \emptyset \;\Rightarrow\; \Pi_{\Delta_{\mathrm{F}}(\phi,\mathfrak{M})} \neq \emptyset,$$

(5.34) becomes true.

Example 5.4.16 Let ϕ contain the L^σ-sentences (3.1) and the sentences

$$\mathrm{prob}(\mathsf{E}\,\mathrm{f}) \geq 0.7 \quad \mathrm{prob}(\mathsf{A}\,\mathrm{f}) \leq 0.5.$$

Feasible models of ϕ may differ both with respect to basic properties of the algebra $\mathfrak{A}'$ generated by the extensions of the predicates HE, A, and E – there are feasible models with $I(\mathsf{A}) = \emptyset$ and others with $I(\mathsf{A}) \neq \emptyset$ for instance – as well as with regard to the statistical measure on $\mathfrak{A}'$. For each statistical base S-structure $\mathfrak{M}$, $\Delta_{\mathrm{F}}(\phi,\mathfrak{M}) \upharpoonright \mathfrak{A}'$ is a closed and convex subset of Δ^N, where $N \leq 8$ is the number of atoms of $\mathfrak{A}'$ in $\mathfrak{M}$. As discussed in section 3.4 (p. 82), depending on μ_{f}, $\pi_{\Delta(\phi,\mathfrak{M})}(\mu_{\mathrm{f}})(I(\mathsf{HE}))$ can have any value in [0.64,0.8]. Hence

$$\phi \mid\!\approx^{(\mathbf{R})} \mathrm{prob}(\mathsf{HE}\,\mathrm{f}) \in [0.64, 0.8].$$

Adding the formulas (3.4) to ϕ, we obtain ϕ' with feasible models $\mathfrak{M}$ in which μ_1 is given by (3.5). If $\mathfrak{M}$ is a real-valued structure we know that

$$\pi_{\Delta(\phi',\mathfrak{M})}(\mu_{\mathrm{f}})(I(\mathsf{HE})) \approx 0.739$$

(cf. (3.16)), so that

$$\phi' \mid\!\approx^{\mathbf{R}} \mathrm{prob}(\mathsf{HE}\,\mathrm{f}) \in [0.7385, 0.7395]. \tag{5.36}$$

The last example gives rise to the question whether in fact (5.36) also holds without the restriction to real-valued probabilities, i.e.

$$\phi' \mid\!\approx \mathrm{prob}(\mathsf{HE}\,\mathrm{f}) \in [0.7385, 0.7395]. \tag{5.37}$$

Previous numerical results (5.23),(5.27), and (5.33) were true both with respect to $\mid\!\approx$ and $\mid\!\approx^{\mathbf{R}}$ because their derivation only relied on Jeffrey's rule, which has been seen (when applicable) to perform cross-entropy minimization in any lrc-field. (5.36), on the other hand, being derived by a numerical optimization process, depends on properties of the numerical behaviour of cross-entropy in **R** that might not be shared by cross-entropy in other lrc-fields.

In fact, it is rather doubtful that (5.36) can be derived on the basis of the axioms LRCF as given in chapter 4. However, it is possible to enforce a far-reaching agreement of the numerical behaviour of general logarithmic functions with that of the logarithm in the reals by adding the axiom schema

$$\forall x \in (0,1] \quad q_n(x) \leq \mathsf{ln}(x) \leq p_n(x) \qquad (n = 1, 2, \ldots) \tag{TAY}$$

with

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

obtained from the Taylor-expansion of the logarithm, to the axiom system LRCF. From TAY it can then be proven that approximate information about the numerical behaviour of ln expressible with rational parameters holds in every model of LRCF $\cup$ TAY, precisely:

$$\begin{aligned} \forall r_1, r_2, s_1, s_2 \in \mathbf{Q} \qquad & \mathbf{R} \models \forall x \in [r_1, r_2] \quad s_1 < \mathsf{ln}(x) < s_2 \\ \Rightarrow \quad & \text{LRCF} \cup \text{TAY} \models \forall x \in [r_1, r_2] \quad s_1 < \mathsf{ln}(x) < s_2. \end{aligned}$$

Similar statements can then be made about cross-entropy and cross-entropy minimization, leading to the result that (5.37) holds when we base the semantics of L^β on models of LRCF $\cup$ TAY. A general equivalence between $\phi \mid\!\approx \psi$ and $\phi \mid\!\approx^{\mathbf{R}} \psi$, however, will be limited to formulas ϕ, ψ that satisfy very strict syntactic restrictions. We do not pursue this issue in detail here.

As yet, all the examples presented used a single event symbol only. In the next example information about two event symbols is represented.

Example 5.4.17 Let Better be an antisymmetric relation on the set of mystery films:

$$\forall v_0 v_1 \; (v_0 \neq v_1 \rightarrow (\mathsf{Better}\, v_0 v_1 \leftrightarrow \neg \mathsf{Better}\, v_1 v_0)), \tag{5.38}$$

and assume that the following formula describes a statistical dependence of the Better predicate on the origin of its arguments

$$[\mathsf{Better}\, v_0 v_1 \mid \neg \mathsf{A}\, v_0 \wedge \mathsf{A}\, v_1]_{(v_0,v_1)} = 0.7. \tag{5.39}$$

Assume that for two films f_0 and f_1 we have made observations bearing on their likely origin:

$$\mathrm{prob}(\mathsf{A}\, \mathrm{f}_0) = 0.8 \quad \mathrm{prob}(\mathsf{A}\, \mathrm{f}_1) = 0.4. \tag{5.40}$$

Let ϕ be the conjunction of (5.38)-(5.40), and $\mathbf{e} := (\mathrm{f}_0, \mathrm{f}_1)$. From (5.38) and (5.39), by the homogeneity of the statistical measure μ_2, we can derive

$$\begin{gathered} [\mathsf{Better}\, v_0 v_1 \mid \mathsf{A}\, v_0 \wedge \neg \mathsf{A}\, v_1]_{(v_0,v_1)} = 0.3 \\ [\mathsf{Better}\, v_0 v_1 \mid v_0 \neq v_1 \wedge \mathsf{A}\, v_0 \wedge \mathsf{A}\, v_1]_{(v_0,v_1)} = 0.5 \\ [\mathsf{Better}\, v_0 v_1 \mid v_0 \neq v_1 \wedge \neg \mathsf{A}\, v_0 \wedge \neg \mathsf{A}\, v_1]_{(v_0,v_1)} = 0.5. \end{gathered} \tag{5.41}$$

For $\mathfrak{M}$ a statistical base S-structure, the set $\Delta_{\rm F}(\phi, \mathfrak{M}) \subset \Delta_{\rm F}\mathfrak{A}_{\mathbf{e}}$ is defined by constraints on the subalgebra $\mathfrak{A}'$ generated by $I(\mathsf{A}) \times M$ and $M \times I(\mathsf{A})$.

The restriction $\Delta_{\rm F}(\phi, \mathfrak{M}) \restriction \mathfrak{A}'$ can be represented as the set of measures of the form

$$\begin{array}{cccc} I(\mathsf{A})^c \times I(\mathsf{A})^c & I(\mathsf{A})^c \times I(\mathsf{A}) & I(\mathsf{A}) \times I(\mathsf{A})^c & I(\mathsf{A}) \times I(\mathsf{A}) \\ r - 0.2 & 0.4 - r & 0.8 - r & r \end{array} \tag{5.42}$$

with $r \in [0.2, 0.4]$. Since $\Delta_{\rm F}(\phi, \mathfrak{M}) \restriction \mathfrak{A}'$ is convex, there is at most one $r_0 \in [0.2, 0.4]$ that defines $\nu_{\mathbf{e}} := \pi_{\Delta_{\rm F}(\phi, \mathfrak{M})}(\mu_{\mathbf{e}} \restriction \mathfrak{A}')$. Then $\nu_{\mathbf{e}}(\mathfrak{M}(\mathsf{Better}\, v_0 v_1)(v_0, v_1))$ is given by Jeffrey's rule as

$$\begin{aligned} &0.5(r_0 - 0.2) + 0.7(0.4 - r_0) + 0.3(0.8 - r_0) + 0.5 r_0 \\ &\qquad = 0.5 \cdot 0.2 + 0.7 \cdot 0.4 + 0.3 \cdot 0.8 = 0.62. \end{aligned}$$

As this turns out to be independent from r_0, we can derive

$$\phi \mid\!\approx^{(\mathbf{R})} \mathrm{prob}(\mathsf{Better}\, \mathrm{f}_0 \mathrm{f}_1) = 0.62$$

without knowing the exact subjective probabilities entailed by ϕ for $\neg\mathsf{A}\, \mathrm{f}_0 \wedge \neg\mathsf{A}\, \mathrm{f}_1, \ldots, \mathsf{A}\, \mathrm{f}_0 \wedge \mathsf{A}\, \mathrm{f}_1$ (see also example 5.5.5 below).

One of the most serious shortcomings of *entropy maximization* as a rule of probabilistic inference from information about a single probability distribution (as opposed to default reasoning about probabilities where information about two distributions of different type is combined) is its representation dependency.

One of the best known examples illustrating this effect is the question of the probability for life on mars: if probabilities are assigned according to the maximum entropy principle (which here is just reduced to the classical principle of indifference) to the two propositions "there is life on mars", and "there is no life on mars", then each of the two alternatives will be assigned probability 0.5. By the same principle, to the three propositions "there is animal life on mars", "there is only plant life on mars", and "there is no life on mars" will be assigned probability 1/3 each, yielding a different probability evaluation as before for the proposition "there is no life on mars". The following example illustrates that a similar effect is avoided in minimum cross-entropy reasoning.

Example 5.4.18 Consider the knowledge base ϕ_1 consisting of the conjunction of the following three statements in the vocabulary $\mathrm{S} = \{\mathsf{A}, \mathsf{B}\}$ with an event symbol e

$$\mathrm{prob}(\mathsf{A}\, \mathrm{e}) \geq 0.4 \qquad [\mathsf{B}\, v \mid \mathsf{A}\, v]_v = 0.3 \qquad [\mathsf{B}\, v \mid \neg\mathsf{A}\, v]_v = 0.1. \tag{5.43}$$

The statistical distribution $\mu_{\rm e}$ on the algebra $\mathfrak{A}_{\rm e}$ generated by the four atoms $I(\mathsf{A})^c \cup I(\mathsf{B})^c, \ldots, I(\mathsf{A}) \cup I(\mathsf{B})$ in a feasible model of ϕ_1 here is only partially specified by the two constraints on the conditional distribution of B. Particularly, the statistical probability $\mu_{\rm e}(I(\mathsf{A}))$ is completely undetermined. Depending on its value we will obtain

$$\begin{array}{ll} \pi_{\Delta_{\rm F}(\phi_1, \mathfrak{M})}(\mu_{\rm e})(I(\mathsf{A})) = 0.4 & \text{if } \mu_{\rm e}(I(\mathsf{A})) \leq 0.4 \\ \pi_{\Delta_{\rm F}(\phi_1, \mathfrak{M})}(\mu_{\rm e})(I(\mathsf{A})) = \mu_{\rm e}(I(\mathsf{A})) & \text{if } \mu_{\rm e}(I(\mathsf{A})) \geq 0.4, \end{array}$$

and correspondingly

$$\begin{array}{ll}\pi_{\Delta_{\mathrm{F}}(\phi_1,\mathfrak{M})}(\mu_{\mathrm{e}})(I(\mathsf{B})) = 0.4\cdot 0.3 + 0.6\cdot 0.1 = 0.18 & \text{if } \mu_{\mathrm{e}}(I(\mathsf{A}))\leq 0.4\\ \pi_{\Delta_{\mathrm{F}}(\phi_1,\mathfrak{M})}(\mu_{\mathrm{e}})(I(\mathsf{B})) = \mu_{\mathrm{e}}(I(\mathsf{A}))\cdot 0.3 + \mu_{\mathrm{e}}(I(\mathsf{A})^c)\cdot 0.1 \in [0.18, 0.3] & \text{if } \mu_{\mathrm{e}}(I(\mathsf{A}))\geq 0.4.\end{array}$$

Hence

$$\phi_1 \approx^{(\mathbf{R})} \mathrm{prob}(\mathsf{B}\,\mathrm{e}) \in [0.18, 0.3].$$

Now consider an encoding ϕ_2 of essentially the same information as in ϕ_1, but using the vocabulary $\mathrm{S}' = \{\mathsf{A}_1, \mathsf{A}_2, \mathsf{B}\}$:

$$\mathrm{prob}(\mathsf{A}_1\,\mathrm{e} \vee \mathsf{A}_2\,\mathrm{e}) \geq 0.4 \quad [\mathsf{B}\,v \mid \mathsf{A}_1\,v \vee \mathsf{A}_2\,v]_v = 0.3 \quad [\mathsf{B}\,v \mid \neg(\mathsf{A}_1\,v \vee \mathsf{A}_2\,v)]_v = 0.1. \qquad (5.44)$$

The algebra $\mathfrak{A}_{\mathrm{e}}$ in a belief S′-structure for e here is generated by the eight atoms $I(\mathsf{A}_1)^c \cap I(\mathsf{A}_2)^c \cap I(\mathsf{B})^c, \ldots, I(\mathsf{A}_1)\cap I(\mathsf{A}_2)\cap I(\mathsf{B})$. The statistical distribution μ_{e} now is even "more underspecified" than before, as it can not be completely determined by eliminating a single "degree of freedom", i.e. by way of fixing $\mu_{\mathrm{e}}(I(\mathsf{A}_1)\cap I(\mathsf{A}_2))$ for instance.

However, whatever statistical distribution μ_{e} is used in a particular feasible model of ϕ_2, cross-entropy minimization with respect to this measure and the single constraint $\mathrm{prob}(\mathsf{A}_1\,\mathrm{e} \vee \mathsf{A}_2\,\mathrm{e}) \geq 0.4$ only has to be performed on the subalgebra consisting of the two atoms $I(\mathsf{A}_1)\cup I(\mathsf{A}_2), (I(\mathsf{A}_1)\cup I(\mathsf{A}_2))^c$, by constraints on which $\Delta_{\mathrm{F}}(\phi_2,\mathfrak{M})$ is determined. Depending on the statistical measure of these sets we again obtain

$$\pi_{\Delta_{\mathrm{F}}(\phi_2,\mathfrak{M})}(\mu_{\mathrm{e}})(I(\mathsf{A}_1)\cup I(\mathsf{A}_2)) \in [0.4, 1],$$

and consequently

$$\pi_{\Delta_{\mathrm{F}}(\phi_2,\mathfrak{M})}(\mu_{\mathrm{e}})(I(\mathsf{B})) \in [0.18, 0.3], \quad \text{i.e.}$$

$$\phi_2 \approx^{(\mathbf{R})} \mathrm{prob}(\mathsf{B}\,\mathrm{e}) \in [0.18, 0.3].$$

This invariance of the default entailment $\approx^{(\mathbf{R})}$ when the underlying algebra $\mathfrak{A}_{\mathrm{e}}$ is replaced by a more fine-grained algebra $\mathfrak{A}'_{\mathrm{e}}$ essentially depends on our not making any default assumptions about the statistical measures μ_n. If, prior to the cross-entropy minimization process, the statistical information was subjected to entropy maximization, for instance, then in the above example we would end up with two different statistical measures on the subalgebras $\{I(\mathsf{A}), I(\mathsf{A})^c\}$ and $\{I(\mathsf{A}_1)\cup I(\mathsf{A}_2), (I(\mathsf{A}_1)\cup I(\mathsf{A}_2))^c\}$, when, respectively, interpreting ϕ_1 and ϕ_2, and, consequently, with two different subjective probabilities for $\mathsf{B}\mathrm{e}$.

5.5 Logical Properties of $\approx$

In this section it is shown how the analytical properties of system- and subset independence (theorem 4.0.18 and 4.0.19) translate into logical properties of the entailment relation $\approx$.

Logical properties that emanate from system independence essentially reflect independence assumptions for events when prior beliefs are given by distinct bodies of information for each event. This is most clearly displayed in corollaries 5.5.3 and 5.5.4 below.

Theorem 5.5.1 (Independence) Let S be a vocabulary, $\mathbf{e}'$ and $\mathbf{e}''$ disjoint tuples of event symbols. $\mathbf{e} := (\mathbf{e}', \mathbf{e}'')$. Let $\phi^\sigma \in \mathrm{L}^\sigma_{\mathrm{S}}$, $\phi^{\beta(\mathbf{e}')} \in \mathrm{L}^\beta_{\mathrm{S},\mathbf{e}'}$, $\phi^{\beta(\mathbf{e}'')} \in \mathrm{L}^\beta_{\mathrm{S},\mathbf{e}''}$. Let

$$\phi :\equiv \phi^\sigma \wedge \phi^{\beta(\mathbf{e}')} \wedge \phi^{\beta(\mathbf{e}'')} \in \mathrm{L}^\beta_{\mathrm{S},\mathbf{e}}.$$

Let $\mathfrak{M}$ be a statistical base S-structure, $\mathfrak{A}' \subseteq \mathfrak{A}_{\mathbf{e}'}$ the finite subalgebra generated by the sets $\mathfrak{M}(\chi(\boldsymbol{v}))$ for terms $\mathrm{prob}(\chi[\mathbf{e}'])$ occurring in $\phi^{\beta(\mathbf{e}')}$, and $\mathfrak{A}'' \subseteq \mathfrak{A}_{\mathbf{e}''}$ the finite subalgebra generated by the sets $\mathfrak{M}(\chi(\boldsymbol{v}))$ for terms $\mathrm{prob}(\chi[\mathbf{e}''])$ occurring in $\phi^{\beta(\mathbf{e}'')}$. We have $\mathfrak{A}_{\mathbf{e}'} \times \mathfrak{A}_{\mathbf{e}''} \subseteq \mathfrak{A}_{\mathbf{e}}$, so that for $\nu \in \Delta_{\mathrm{F}}\mathfrak{A}_{\mathbf{e}}$ we can define

$$\begin{aligned}\nu' &:= (\nu \restriction (\mathfrak{A}_{\mathbf{e}'} \times \mathfrak{A}_{\mathbf{e}''}))\ |_1\ \mathfrak{A}_{\mathbf{e}'},\\ \nu'' &:= (\nu \restriction (\mathfrak{A}_{\mathbf{e}'} \times \mathfrak{A}_{\mathbf{e}''}))\ |_2\ \mathfrak{A}_{\mathbf{e}''}.\end{aligned}$$

Then $(\mathfrak{M}, \nu_{\mathbf{e}}) \mathrel{|\!\!\approx} \phi$ iff the following four conditions are satisfied:

$$\begin{array}{ll}\text{(i)} & (\mathfrak{M}, \nu_{\mathbf{e}}) \models_\beta \phi\\ \text{(ii)} & (\mathfrak{M}, \nu'_{\mathbf{e}}) \mathrel{|\!\!\approx} \phi^\sigma \wedge \phi^{\beta(\mathbf{e}')}\\ \text{(iii)} & (\mathfrak{M}, \nu''_{\mathbf{e}}) \mathrel{|\!\!\approx} \phi^\sigma \wedge \phi^{\beta(\mathbf{e}'')}\\ \text{(iv)} & \nu_{\mathbf{e}} = \mathcal{J}(\nu'_{\mathbf{e}} \restriction \mathfrak{A}' \otimes \nu''_{\mathbf{e}} \restriction \mathfrak{A}'', \mu_{\mathbf{e}}, \mathfrak{A}_{\mathbf{e}}).\end{array}$$

Proof: Let $(\mathfrak{M},\nu_{\mathbf{e}})$ be a belief structure. By the definition of $\models_\beta$ then

$$\begin{aligned}(\mathfrak{M}, \nu_{\mathbf{e}}) \models_\beta \phi \ &\Leftrightarrow (\mathfrak{M}, \nu_{\mathbf{e}}) \models_\beta \phi^\sigma \wedge \phi^{\beta(\mathbf{e}')} \ \text{ and } \ (\mathfrak{M}, \nu_{\mathbf{e}}) \models_\beta \phi^\sigma \wedge \phi^{\beta(\mathbf{e}'')}\\ &\Leftrightarrow (\mathfrak{M}, \nu'_{\mathbf{e}}) \models_\beta \phi^\sigma \wedge \phi^{\beta(\mathbf{e}')} \ \text{ and } \ (\mathfrak{M}, \nu''_{\mathbf{e}}) \models_\beta \phi^\sigma \wedge \phi^{\beta(\mathbf{e}'')}.\end{aligned} \tag{5.45}$$

From (5.45) we obtain a representation of $\Delta(\phi, \mathfrak{M})$:

$$\begin{aligned}\Delta_{\mathrm{F}}(\phi, \mathfrak{M}) &= \{\nu \in \Delta_{\mathrm{F}}\mathfrak{A}_{\mathbf{e}} \mid \nu' \in \Delta_{\mathrm{F}}(\phi^\sigma \wedge \phi^{\beta(\mathbf{e}')}, \mathfrak{M}) \text{ and } \nu'' \in \Delta_{\mathrm{F}}(\phi^\sigma \wedge \phi^{\beta(\mathbf{e}'')}, \mathfrak{M})\}\\ &= \{\nu \in \Delta_{\mathrm{F}}\mathfrak{A}_{\mathbf{e}} \mid \nu' \restriction \mathfrak{A}' \in \Delta_{\mathrm{F}}(\phi^\sigma \wedge \phi^{\beta(\mathbf{e}')}, \mathfrak{M}) \restriction \mathfrak{A}' \text{ and}\\ &\qquad\qquad \nu'' \restriction \mathfrak{A}'' \in \Delta_{\mathrm{F}}(\phi^\sigma \wedge \phi^{\beta(\mathbf{e}'')}, \mathfrak{M}) \restriction \mathfrak{A}''\}.\end{aligned}$$

By theorem 4.0.18 then

$$\begin{aligned}&\Pi_{\Delta_{\mathrm{F}}(\phi,\mathfrak{M}) \restriction \mathfrak{A}' \times \mathfrak{A}''}(\mu_{\mathbf{e}} \restriction (\mathfrak{A}' \times \mathfrak{A}'')) =\\ &\qquad \Pi_{\Delta_{\mathrm{F}}(\phi^\sigma \wedge \phi^{\beta(\mathbf{e}')}, \mathfrak{M}) \restriction \mathfrak{A}'}(\mu_{\mathbf{e}'} \restriction \mathfrak{A}') \otimes \Pi_{\Delta_{\mathrm{F}}(\phi^\sigma \wedge \phi^{\beta(\mathbf{e}'')}, \mathfrak{M}) \restriction \mathfrak{A}''}(\mu_{\mathbf{e}''} \restriction \mathfrak{A}'').\end{aligned} \tag{5.46}$$

To now prove the left to right direction of the equivalence stated in the theorem, assume that $(\mathfrak{M}, \nu_{\mathbf{e}}) \mathrel{|\!\!\approx} \phi$. This immediately implies (i), and by (5.45) the feasibility assertions contained in (ii) and (iii). Since $\Delta_{\mathrm{F}}(\phi, \mathfrak{M})$ is defined by constraints on $\mathfrak{A}' \times \mathfrak{A}''$, we have

$$\nu_{\mathbf{e}} \restriction (\mathfrak{A}' \times \mathfrak{A}'') \in \Pi_{\Delta_{\mathrm{F}}(\phi,\mathfrak{M}) \restriction \mathfrak{A}' \times \mathfrak{A}''}(\mu_{\mathbf{e}} \restriction (\mathfrak{A}' \times \mathfrak{A}'')),$$

and hence with (5.46)

$$\begin{aligned}\nu'_{\mathbf{e}} \restriction \mathfrak{A}' &= (\nu_{\mathbf{e}} \restriction (\mathfrak{A}' \times \mathfrak{A}''))\ |_1\ \mathfrak{A}' \in \Pi_{\Delta_{\mathrm{F}}(\phi^\sigma \wedge \phi^{\beta(\mathbf{e}')}, \mathfrak{M}) \restriction \mathfrak{A}'}(\mu_{\mathbf{e}'} \restriction \mathfrak{A}'),\\ \nu''_{\mathbf{e}} \restriction \mathfrak{A}'' &= (\nu_{\mathbf{e}} \restriction (\mathfrak{A}' \times \mathfrak{A}''))\ |_2\ \mathfrak{A}'' \in \Pi_{\Delta_{\mathrm{F}}(\phi^\sigma \wedge \phi^{\beta(\mathbf{e}'')}, \mathfrak{M}) \restriction \mathfrak{A}''}(\mu_{\mathbf{e}''} \restriction \mathfrak{A}''),\end{aligned} \tag{5.47}$$

and $\nu_{\mathbf{e}}\restriction(\mathfrak{A}' \times \mathfrak{A}'') = \nu'_{\mathbf{e}}\restriction\mathfrak{A}' \otimes \nu''_{\mathbf{e}}\restriction\mathfrak{A}''$, which proves (iv).

To also prove (ii), it must be shown that

$$\nu'_{\mathbf{e}} = \mathcal{J}(\nu'_{\mathbf{e}}\restriction\mathfrak{A}', \mu_{\mathbf{e}'}, \mathfrak{A}_{\mathbf{e}'}), \tag{5.48}$$

because then, with (5.47), it follows that

$$\nu'_{\mathbf{e}} \in \Pi_{\Delta(\phi^{\sigma}\wedge\phi^{\beta}(\mathbf{e}'),\mathfrak{M})}(\mu_{\mathbf{e}'}).$$

For the proof of (5.48) let $\{A'_1, \ldots, A'_p\}$, $\{A''_1, \ldots, A''_q\}$ be the atoms of $\mathfrak{A}'$ and $\mathfrak{A}''$ respectively. The set of atoms of $\mathfrak{A}' \times \mathfrak{A}''$ then is

$$\{A_{ij} := A'_i \times A''_j \mid i \in \{1, \ldots, p\},\ j \in \{1, \ldots, q\}\}.$$

Let $A' \in \mathfrak{A}_{\mathbf{e}'}$. First note that, due to the product-property of $\mu_{\mathbf{e}}$, for all $A'' \in \mathfrak{A}_{\mathbf{e}''}$

$$\begin{aligned} \mu_{\mathbf{e}}(A' \times A'' \mid A_{ij}) &= \frac{\mu_{\mathbf{e}}((A' \times A'') \cap (A'_i \times A''_j))}{\mu_{\mathbf{e}}(A'_i \times A''_j)} \\ &= \frac{\mu_{\mathbf{e}'}(A' \cap A'_i)\mu_{\mathbf{e}''}(A'' \cap A''_j)}{\mu_{\mathbf{e}'}(A'_i)\mu_{\mathbf{e}''}(A''_j)} \\ &= \mu_{\mathbf{e}'}(A' \mid A'_i)\mu_{\mathbf{e}''}(A'' \mid A''_j). \end{aligned} \tag{5.49}$$

Thus

$$\begin{aligned} \nu'_{\mathbf{e}}(A') &= \nu_{\mathbf{e}}(A' \times M^{|\mathbf{e}''|}) \\ &= \textstyle\sum_{ij}(\nu'_{\mathbf{e}}\restriction\mathfrak{A}' \otimes \nu''_{\mathbf{e}}\restriction\mathfrak{A}'')(A_{ij})\mu_{\mathbf{e}}(A' \times M^{|\mathbf{e}''|} \mid A_{ij}) && \text{by (iv)} \\ &= \textstyle\sum_{ij} \nu'_{\mathbf{e}}\restriction\mathfrak{A}'(A'_i)\nu''_{\mathbf{e}}\restriction\mathfrak{A}''(A''_j)\mu_{\mathbf{e}'}(A' \mid A'_i) && \text{by (5.49)} \\ &= \textstyle\sum_i \nu'_{\mathbf{e}}\restriction\mathfrak{A}'(A'_i)\mu_{\mathbf{e}'}(A' \mid A'_i) \\ &= \mathcal{J}(\nu'_{\mathbf{e}}\restriction\mathfrak{A}', \mu_{\mathbf{e}'}, \mathfrak{A}_{\mathbf{e}'})(A'). \end{aligned}$$

This proves (5.48), and hence (ii). The proof of (iii) is identical.

That conversely (i)-(iv) implies $(\mathfrak{M}, \nu_{\mathbf{e}}) \approx \phi$ follows easily with (5.46): from (ii) and (iii) we immediately obtain (5.47), which with (5.46) implies that

$$\mathcal{J}(\nu'_{\mathbf{e}}\restriction\mathfrak{A}' \otimes \nu''_{\mathbf{e}}\restriction\mathfrak{A}'', \mu_{\mathbf{e}}, \mathfrak{A}_{\mathbf{e}}) \in \Pi_{\Delta(\phi,\mathfrak{M})}(\mu_{\mathbf{e}}).$$

(i) and (iv) together then just state that $(\mathfrak{M}, \nu_{\mathbf{e}}) \approx \phi$. □

Corollary 5.5.2 Let ϕ be as in theorem 5.5.1; $(\mathfrak{M}, \nu_{\mathbf{e}}) \approx \phi$. Then $\nu_{\mathbf{e}}$ is an extension of $\nu'_{\mathbf{e}} \otimes \nu''_{\mathbf{e}} \in \Delta_{\mathrm{F}}(\mathfrak{A}_{\mathbf{e}'} \times \mathfrak{A}_{\mathbf{e}''})$.

Proof: Let A'_i, A''_j, A_{ij} $(i = 1, \ldots, p;\ j = 1, \ldots, q)$ be as in the proof of theorem 5.5.1. Let

$A \in \mathfrak{A}_{\mathbf{e}'} \times \mathfrak{A}_{\mathbf{e}''}$. It suffices to consider the case $A = A' \times A''$ with $A' \in \mathfrak{A}_{\mathbf{e}'}, A'' \in \mathfrak{A}_{\mathbf{e}''}$. Then

$$\begin{aligned}
\nu_{\mathbf{e}}(A' \times A'') &= \mathcal{J}(\nu'_{\mathbf{e}} \upharpoonright \mathfrak{A}' \otimes \nu''_{\mathbf{e}} \upharpoonright \mathfrak{A}'', \mu_{\mathbf{e}}, \mathfrak{A}_{\mathbf{e}})(A' \times A'') && \text{by (iv)} \\
&= \sum_{ij} (\nu'_{\mathbf{e}} \upharpoonright \mathfrak{A}' \otimes \nu''_{\mathbf{e}} \upharpoonright \mathfrak{A}'')(A_{ij}) \mu_{\mathbf{e}}(A' \times A'' \mid A_{ij}) \\
&= \sum_{ij} \nu'_{\mathbf{e}} \upharpoonright \mathfrak{A}'(A'_i) \nu''_{\mathbf{e}} \upharpoonright \mathfrak{A}''(A''_j) \mu_{\mathbf{e}'}(A' \mid A'_i) \mu_{\mathbf{e}''}(A'' \mid A''_j) && \text{by (5.49)} \\
&= \mathcal{J}(\nu'_{\mathbf{e}} \upharpoonright \mathfrak{A}', \mu_{\mathbf{e}'}, \mathfrak{A}_{\mathbf{e}'})(A') \mathcal{J}(\nu''_{\mathbf{e}} \upharpoonright \mathfrak{A}'', \mu_{\mathbf{e}''}, \mathfrak{A}_{\mathbf{e}''})(A'') \\
&= (\nu'_{\mathbf{e}} \otimes \nu''_{\mathbf{e}})(A' \times A'') && \text{by (5.48).}
\end{aligned}$$

□

Corollary 5.5.3 Let ϕ be as in theorem 5.5.1, $\chi'(\boldsymbol{v}'), \chi''(\boldsymbol{v}'') \in \mathrm{L}^\sigma$ with $|\boldsymbol{v}'| = |\mathbf{e}'|$ and $|\boldsymbol{v}''| = |\mathbf{e}''|$. Then

$$\phi \mathrel{|}\!\!\approx^{(\mathbf{R})} \mathrm{prob}(\chi'[\mathbf{e}'] \wedge \chi''[\mathbf{e}'']) = \mathrm{prob}(\chi'[\mathbf{e}'])\mathrm{prob}(\chi''[\mathbf{e}'']).$$

Proof: Immediate from corollary 5.5.2. □

Corollary 5.5.4 Let ϕ, χ', χ'' be as in the preceding corollary, $r', r'' \in \mathbf{Q}$. Then, with $\sim$ any relation from $\{\leq, <, \geq, >, =\}$:

$$\begin{gathered}
\phi^\sigma \wedge \phi^{\beta(\mathbf{e}')} \mathrel{|}\!\!\approx^{(\mathbf{R})} \mathrm{prob}(\chi'[\mathbf{e}']) \sim r' \quad \wedge \quad \phi^\sigma \wedge \phi^{\beta(\mathbf{e}'')} \mathrel{|}\!\!\approx^{(\mathbf{R})} \mathrm{prob}(\chi''[\mathbf{e}'']) \sim r'' \\
\Rightarrow \phi \mathrel{|}\!\!\approx^{(\mathbf{R})} \mathrm{prob}(\chi'[\mathbf{e}'] \wedge \chi''[\mathbf{e}'']) \sim r'r''.
\end{gathered}$$

Proof: Suppose that $(\mathfrak{M}, \nu_{\mathbf{e}}) \mathrel{|}\!\!\approx \phi$. By theorem 5.5.1 then $(\mathfrak{M}, \nu'_{\mathbf{e}}) \mathrel{|}\!\!\approx \phi^\sigma \wedge \phi^{\beta(\mathbf{e}')}$ and $(\mathfrak{M}, \nu''_{\mathbf{e}}) \mathrel{|}\!\!\approx \phi^\sigma \wedge \phi^{\beta(\mathbf{e}'')}$. Hence

$$\nu'_{\mathbf{e}}(\mathfrak{M}(\chi'(\boldsymbol{v}'))) \sim r' \quad \text{and} \quad \nu''_{\mathbf{e}}(\mathfrak{M}(\chi''(\boldsymbol{v}''))) \sim r''.$$

By corollary 5.5.2 therefore

$$\nu_{\mathbf{e}}(\mathfrak{M}((\chi' \wedge \chi'')(\boldsymbol{v}', \boldsymbol{v}''))) \sim r'r''.$$

□

Example 5.5.5 Reconsider ϕ given in example 5.4.17. ϕ is of the form dealt with in theorem 5.5.1. From corollary 5.5.2 it therefore follows that for a statistical base S-structure $\mathfrak{M}$, and $\nu_{\mathbf{e}} = \pi_{\Delta_{\mathrm{F}}(\phi, \mathfrak{M})}(\mu_{\mathbf{e}})$:

$$\nu_{\mathbf{e}}(I(\mathsf{A}) \times I(\mathsf{A})) = \nu'_{\mathbf{e}}(I(\mathsf{A}))\nu''_{\mathbf{e}}(I(\mathsf{A})) = 0.8 \cdot 0.4 = 0.32.$$

Hence $\nu_{\mathbf{e}}$ is defined by the parameter $r_0 = 0.32$ in (5.42), and

$$\phi \mathrel{|}\!\!\approx^{(\mathbf{R})} \mathrm{prob}(\mathsf{A}\,\mathrm{f}_0 \wedge \mathsf{A}\,\mathrm{f}_1) = 0.32.$$

The next theorem translates the subset independence property (theorem 4.0.19) into a property of $\mathscr{L}^{\beta}_{\text{def}}$. The logical character of this property is displayed by corollary 5.5.7 below.

Theorem 5.5.6 (Reasoning by Cases) Let S be a vocabulary, $\mathbf{e}$ a tuple of event symbols. Let $\phi^{\sigma}, \chi_1(\boldsymbol{v}), \ldots, \chi_L(\boldsymbol{v}) \in \mathrm{L}^{g}_{\mathrm{S}}$ with $|\boldsymbol{v}|=|\mathbf{e}|$ such that

$$\phi^{\sigma} \models_{\beta} \forall \boldsymbol{v}(\chi_1(\boldsymbol{v}) \dot{\vee} \ldots \dot{\vee} \chi_L(\boldsymbol{v})).$$

Let $\bar{\phi} \in \mathrm{L}^{\beta}_{\mathrm{S},\mathbf{e}}$ only contain subjective probability terms of the form $\mathrm{prob}(\chi_h[\mathbf{e}])$ ($h \in \{1, \ldots, L\}$). For $h = 1, \ldots, L$ let $\phi^h \in \mathrm{L}^{\beta}_{\mathrm{S},\mathbf{e}}$ only contain subjective probability terms of the form

$$\mathrm{prob}(\psi[\mathbf{e}] \mid (\chi_h \wedge \psi')[\mathbf{e}]) \quad (\psi(\boldsymbol{v}), \psi'(\boldsymbol{v}) \in \mathrm{L}^{g}_{\mathrm{S}})$$

(more precisely: ϕ^h can be abbreviated in such a way as to only contain terms of this form, cf. p. 98).

Let

$$\phi :\equiv \phi^{\sigma} \wedge \bar{\phi} \wedge \phi^1 \wedge \ldots \wedge \phi^L.$$

Let $\mathfrak{M}$ be a statistical base S-structure. For $\nu \in \Delta_{\mathrm{F}}\mathfrak{A}_{\mathbf{e}}$ with $\nu(\mathfrak{M}(\chi_h(\boldsymbol{v}))) > 0$ let ν^h denote the conditional distribution of ν on $\mathfrak{M}(\chi_h(\boldsymbol{v}))$.

If $(\mathfrak{M}, \nu_{\mathbf{e}}) \mathrel{|\!\!\approx} \phi$ then for all $h \in \{1, \ldots, L\}$ with $\nu_{\mathbf{e}}(\mathfrak{M}(\chi_h(\boldsymbol{v})) > 0$ there exists $\tilde{\nu}_{\mathbf{e}} \in \Delta_{\mathrm{F}}\mathfrak{A}_{\mathbf{e}}$ with

$$\tilde{\nu}^h_{\mathbf{e}} = \nu^h_{\mathbf{e}} \ \text{ and } \ (\mathfrak{M}, \tilde{\nu}_{\mathbf{e}}) \mathrel{|\!\!\approx} \phi^{\sigma} \wedge \phi^h \wedge \mathrm{prob}(\chi_h[\mathbf{e}]) = 1.$$

Proof: Let $(\mathfrak{M}, \nu_{\mathbf{e}}) \mathrel{|\!\!\approx} \phi$. Let $\mathfrak{A} \subseteq \mathfrak{A}_{\mathbf{e}}$ be the finite subalgebra generated by the sets $\mathfrak{M}(\chi_h(\boldsymbol{v}))$, $\mathfrak{M}(\psi(\boldsymbol{v}))$, and $\mathfrak{M}(\psi'(\boldsymbol{v}))$ with $\mathrm{prob}(\psi[\mathbf{e}] \mid (\chi_h \wedge \psi')[\mathbf{e}])$ a term in ϕ^h ($h = 1, \ldots, L$). Let $\bar{\mathfrak{A}} \subseteq \mathfrak{A}$ be the algebra generated by the sets $\mathfrak{M}(\chi_h(\boldsymbol{v}))$ alone, and $\mathfrak{A}^h$ the relative algebra of $\mathfrak{A}$ with respect to $\mathfrak{M}(\chi_h(\boldsymbol{v}))$ (cf. theorem 4.0.19).

Then

$$\Delta_{\mathrm{F}}(\phi, \mathfrak{M}) = \Delta_{\mathrm{F}}(\bar{\phi}, \mathfrak{M}) \cap \bigcap_{h=1}^{L} \Delta_{\mathrm{F}}(\phi^h, \mathfrak{M})$$

with $\Delta_{\mathrm{F}}(\bar{\phi}, \mathfrak{M})$ a set of constraints on the distribution on $\bar{\mathfrak{A}}$, and $\Delta_{\mathrm{F}}(\phi^h, \mathfrak{M})$ a set of constraints on the conditional distribution on $\mathfrak{A}^h$ ($h = 1, \ldots, L$).

Making restrictions to the finite algebras $\mathfrak{A}, \bar{\mathfrak{A}}, \mathfrak{A}^h$, we are in the situation described in theorem 4.0.19 with

$$\begin{aligned}
\bar{J} &= \{\nu \in \Delta_{\mathrm{F}}(\bar{\phi}, \mathfrak{M}) \upharpoonright \mathfrak{A} \mid \bar{\nu} \in \bar{J}^*\} \ \text{ with } \ \bar{J}^* = \{\bar{\nu} \in \Delta_{\mathrm{F}}\bar{\mathfrak{A}} \mid \nu \in \Delta_{\mathrm{F}}(\bar{\phi}, \mathfrak{M})\} \\
J_h &= \{\nu \in \Delta_{\mathrm{F}}(\phi_h, \mathfrak{M}) \upharpoonright \mathfrak{A} \mid \nu(\mathfrak{M}(\chi_h(\boldsymbol{v}))) = 0 \ \vee \ \nu^h \in J^*_h\} \ \text{ with} \\
&\qquad J^*_h = \{\nu^h \in \Delta_{\mathrm{F}}\mathfrak{A}^h \mid \nu(\mathfrak{M}(\chi_h(\boldsymbol{v}))) > 0 \ \wedge \ \nu \in \Delta_{\mathrm{F}}(\phi^h, \mathfrak{M})\}.
\end{aligned}$$

Now let $h \in \{1, \ldots, L\}$ with $\nu_{\mathbf{e}}(\mathfrak{M}(\chi_h(\boldsymbol{v}))) > 0$ (and consequently $\mu_{\mathbf{e}}(\mathfrak{M}(\chi_h(\boldsymbol{v}))) > 0$). By theorem 4.0.19 then

$$\nu^h_{\mathbf{e}} \upharpoonright \mathfrak{A}^h \in \Pi_{J^*_h}(\mu^h_{\mathbf{e}} \upharpoonright \mathfrak{A}^h). \tag{5.50}$$

Now define $\tilde{\nu}_{\mathbf{e}} \in \Delta_{\mathrm{F}}\mathfrak{A}_{\mathbf{e}}$ by

$$\tilde{\nu}_{\mathbf{e}}(\mathfrak{M}(\chi_h(\boldsymbol{v}))) = 1, \tag{5.51}$$

$$\tilde{\nu}^h_{\mathbf{e}} = \nu^h_{\mathbf{e}}. \tag{5.52}$$

We have to show that with $\gamma :\equiv \phi^\sigma \wedge \phi^h \wedge \mathrm{prob}(\chi_h[\mathbf{e}]) = 1$,

$$\tilde{\nu}_{\mathbf{e}} \in \Pi_{\Delta_{\mathrm{F}}(\gamma,\mathfrak{M})}(\mu_{\mathbf{e}}). \tag{5.53}$$

Clearly, $\tilde{\nu}_{\mathbf{e}} \in \Delta_{\mathrm{F}}(\gamma, \mathfrak{M})$, so that it remains to be shown that $\tilde{\nu}_{\mathbf{e}}$ minimizes $CE^{\mathrm{F}}(\cdot, \mu_{\mathbf{e}})$ within this set.

By lemma 4.0.16, for all $\nu \in \Delta_{\mathrm{F}}(\gamma, \mathfrak{M})$

$$CE^{\mathrm{F}}(\nu \upharpoonright \mathfrak{A}, \mu_{\mathbf{e}} \upharpoonright \mathfrak{A}) = \mathsf{ln}^{\mathrm{F}}(1/\mu_{\mathbf{e}}(\mathfrak{M}(\chi_h(\boldsymbol{v})))) + CE^{\mathrm{F}}(\nu^h \upharpoonright \mathfrak{A}^h, \mu^h_{\mathbf{e}} \upharpoonright \mathfrak{A}^h). \tag{5.54}$$

Since

$$\Delta_{\mathrm{F}}(\gamma, \mathfrak{M}) \subseteq \{\nu \mid \nu^h \upharpoonright \mathfrak{A}^h \in J^*_h\},$$

(5.54) will certainly be minimal when $\nu^h \upharpoonright \mathfrak{A}^h \in \Pi_{J^*_h}(\mu^h_{\mathbf{e}} \upharpoonright \mathfrak{A}^h)$. This, by (5.50) and (5.52), is true for $\tilde{\nu}^h_{\mathbf{e}}$, so that

$$\tilde{\nu}_{\mathbf{e}} \upharpoonright \mathfrak{A} \in \Pi_{\Delta_{\mathrm{F}}(\gamma,\mathfrak{M}) \upharpoonright \mathfrak{A}}(\mu_{\mathbf{e}} \upharpoonright \mathfrak{A}).$$

Finally, the restriction to $\mathfrak{A}$ can be dropped, because from $\nu_{\mathbf{e}} = \mathcal{J}(\nu_{\mathbf{e}} \upharpoonright \mathfrak{A}, \mu_{\mathbf{e}}, \mathfrak{A}_{\mathbf{e}})$ it follows from the definition of $\tilde{\nu}_{\mathbf{e}}$ that $\tilde{\nu}_{\mathbf{e}} = \mathcal{J}(\tilde{\nu}_{\mathbf{e}} \upharpoonright \mathfrak{A}, \mu_{\mathbf{e}}, \mathfrak{A}_{\mathbf{e}})$. □

Corollary 5.5.7 Let ϕ be as in theorem 5.5.6. Let $h \in \{1, \ldots, L\}$ such that

$$\phi^\sigma \wedge \bar{\phi} \models_\beta \mathrm{prob}(\chi_h[\mathbf{e}]) > 0.$$

Let $\zeta \in \mathrm{L}^\beta_{\mathrm{S},\mathbf{e}}$ only contains subjective probability terms or the form $\mathrm{prob}(\psi[\mathbf{e}] \mid (\chi_h \wedge \psi')[\mathbf{e}])$.

If

$$\phi^\sigma \wedge \phi^h \wedge \mathrm{prob}(\chi_h[\mathbf{e}]) = 1 \mathrel{|\!\!\approx^{(\mathbf{R})}} \zeta,$$

then

$$\phi \mathrel{|\!\!\approx^{(\mathbf{R})}} \zeta.$$

Proof: Let $(\mathfrak{M}, \nu_{\mathbf{e}}) \mathrel{|\!\!\approx} \phi$. Then $\nu_{\mathbf{e}}(\mathfrak{M}(\chi(\boldsymbol{v}))) > 0$, and with $\tilde{\nu}_{\mathbf{e}}$ as given by theorem 5.5.6, $(\mathfrak{M}, \tilde{\nu}_{\mathbf{e}}) \mathrel{|\!\!\approx} \phi^\sigma \wedge \phi^h \wedge \mathrm{prob}(\chi_h[\mathbf{e}]) = 1$. Then $(\mathfrak{M}, \tilde{\nu}_{\mathbf{e}}) \models_\beta \zeta$, and because $\nu^h_{\mathbf{e}} = \tilde{\nu}^h_{\mathbf{e}}$, $(\mathfrak{M}, \nu_{\mathbf{e}}) \models_\beta \zeta$. □

From this corollary it becomes clear why theorem 5.5.6 is understood to be essentially a result concerning the possibility to use reasoning by cases for the inference relation $\mathrel{|\!\!\approx^{(\mathbf{R})}}$. When our knowledge base is of such a structure that it distinguishes several mutually exclusive hypotheses for the events $\mathbf{e}$, and, apart from making general statements about the relative likelihoods for the individual hypotheses to be true, only contains a collection of belief statements each of which is conditioned on one of the hypothesis, then we may make valid default inferences about subjective probabilities conditioned on one of the hypothesis by ignoring all the beliefs pertaining to other hypotheses.

Note, though, that the converse of the corollary by no means is true: from $\phi \mid\!\approx^{(\mathbf{R})} \zeta$ it does not follow that $\phi^\sigma \wedge \phi^h \wedge \mathrm{prob}(\chi_h[\mathbf{e}]) = 1 \mid\!\approx^{(\mathbf{R})} \zeta$ because a default model of $\phi^\sigma \wedge \phi^h \wedge \mathrm{prob}(\chi_h[\mathbf{e}]) = 1$ usually will not even be a feasible model of ϕ.

Example 5.5.8 Consider the properties of a mystery film to be American (A) or French (F), to have a happy end (HE), and to have a gangster as the hero (G). Some deterministic and statistical information about these properties is encoded in the following sentences.

$$\forall v(\neg(\mathsf{A}v \wedge \mathsf{F}v)) \tag{5.55}$$

$$[\mathsf{G}v \mid \mathsf{A}v]_v = 0.1 \tag{5.56}$$

$$[\mathsf{G}v \mid \mathsf{F}v]_v = 0.6 \tag{5.57}$$

$$[\mathsf{G}v \mid \mathsf{F}v \wedge \mathsf{HE}v]_v = 0.3 \tag{5.58}$$

$$[\mathsf{G}v \mid \mathsf{F}v \wedge \neg\mathsf{HE}v]_v = 0.8 \tag{5.59}$$

Suppose some information about a mystery film f has been obtained, which induces the beliefs

$$\mathrm{prob}(\mathsf{A}\,\mathrm{f}) \geq 0.01 \tag{5.60}$$

$$\mathrm{prob}(\mathsf{F}\,\mathrm{f}) \geq 0.01 \tag{5.61}$$

$$\mathrm{prob}(\mathsf{G}\,\mathrm{f} \mid \mathsf{A}\,\mathrm{f}) \geq 0.4 \tag{5.62}$$

$$\mathrm{prob}(\mathsf{HE}\,\mathrm{f} \mid \mathsf{A}\,\mathrm{f}) \geq 0.8 \tag{5.63}$$

The conjunction ϕ of these sentences is of the form described in theorem 5.5.6 with $\chi_1(v) \equiv \mathsf{A}v$, $\chi_2(v) \equiv \mathsf{F}v$, $\chi_3(v) \equiv \neg(\mathsf{F}v \vee \mathsf{A}v)$, ϕ^σ the conjunction of (5.55)-(5.59), $\bar{\phi}$ the conjunction of (5.60) and (5.61), and ϕ^1 the conjunction of (5.62) and (5.63).

Since by Jeffrey's rule

$$\phi^\sigma \wedge \mathrm{prob}(\mathsf{F}\,\mathrm{f}) = 1 \mid\!\approx^{(\mathbf{R})} \mathrm{prob}(\mathsf{G}\,\mathrm{f}) = \mathrm{prob}(\mathsf{G}\,\mathrm{f} \mid \mathsf{F}\,\mathrm{f}) = 0.6,$$

from corollary 5.5.7 we then know that the same result follows from the complete knowledge base:

$$\phi \mid\!\approx^{(\mathbf{R})} \mathrm{prob}(\mathsf{G}\,\mathrm{f} \mid \mathsf{F}\,\mathrm{f}) = 0.6.$$

With the additional prior belief

$$\phi^2 :\equiv \mathrm{prob}(\mathsf{HE}\,\mathrm{f} \mid \mathsf{F}\,\mathrm{f}) = 0.7$$

we obtain

$$\phi^\sigma \wedge \phi^2 \wedge \mathrm{prob}(\mathsf{F}\,\mathrm{f}) = 1 \mid\!\approx^{(\mathbf{R})} \mathrm{prob}(\mathsf{G}\,\mathrm{f} \mid \mathsf{F}\,\mathrm{f}) = 0.7 \cdot 0.3 + 0.3 \cdot 0.8 = 0.45,$$

and hence

$$\phi \wedge \phi^2 \mid\!\approx^{(\mathbf{R})} \mathrm{prob}(\mathsf{G}\,\mathrm{f} \mid \mathsf{F}\,\mathrm{f}) = 0.45.$$

Shore and Johnson ([1980],[1983]) have postulated the subset- and system independence properties as two axioms that should be satisfied by a procedure for updating an estimate for a probability distribution (in Shore and Johnson's context thought of as characterizing a physical system, rather than an epistemic state) when new information about the actual distribution is obtained. Information they always assumed to consist of a closed and convex subset of probability distributions the actual distribution is known to belong to. Conversely, every such closed and convex set constitutes possible information in the sense of Shore and Johnson.

Assuming closed and convex constraint sets, it is reasonable to require an updating procedure to always yield a unique solution as the new estimated probability distribution. This *uniqueness* property is a further axiom that Shore and Johnson postulate. Also they require that the updating procedure is *invariant*, which for the case of distributions on finite spaces just means that it is independent from the order imposed on the atoms of the space, and that the procedure does not change the prior estimate when the new information is trivial, i.e. consists of the set of all probability measures.

From these five axioms Shore and Johnson derive the minimum cross-entropy principle: they show that if an updating procedure is given by minimizing a differentiable function of pairs of probability measures, and satisfies the axioms, then it must be equivalent to cross-entropy minimization.

Using this result, one may obtain an axiomatic derivation of the minimum cross-entropy principle for the default semantics of L^β with respect to real-valued probabilities: if the default entailment relation $\phi \mathrel{|\!\!\approx^{\mathbf{R}}} \ldots$ by a suitable set of axioms is prescribed to behave in a certain way for $\phi \in \mathrm{L}^\beta$ of specific structures, then it follows that the selection of a default measure $\nu_{\mathbf{e}} \in \Delta(\phi, \mathfrak{M})$ (assuming it is performed by minimizing some smooth "distance"-function) has to follow the minimum cross-entropy principle.

The axioms one would lay down for $\mathrel{|\!\!\approx^{\mathbf{R}}}$, of course, would closely correspond to the original axioms of Shore and Johnson, with corollary 5.5.4 and corollary 5.5.7 the two key axioms corresponding to system- and subset independence.

From the axioms for $\mathrel{|\!\!\approx^{\mathbf{R}}}$ then a collection of rules for the selection

$$\mu, \Delta(\phi, \mathfrak{M}) \mapsto \Pi_{\Delta(\phi,\mathfrak{M})}(\mu)$$

for specific instances of μ and $\Delta(\phi, \mathfrak{M})$ can be obtained. These rules will then have to be shown to capture as much of the contents of the original axioms of Shore and Johnson as is actually needed for deriving the minimum cross-entropy principle. The original axioms, in this way, will not be recovered in full, mainly because these contain a quantification over all closed and convex sets of probability measures as admissible constraint sets, while the rules obtained from axioms for $\mathrel{|\!\!\approx^{\mathbf{R}}}$ will only refer to sets $\Delta(\phi, \mathfrak{M})$ definable by some $\phi \in \mathrm{L}^\beta$.

However, the proof of Shore and Johnson makes no use of the existence of constraint sets other than can also be defined in L^β, so that the proof may be carried out on the basis of the rules obtained from axioms for $\mathrel{|\!\!\approx^{\mathbf{R}}}$.

A result closely related to the Shore and Johnson derivation of the minimum cross-entropy principle has been shown by Paris and Vencovská ([1990]). From a set of seven axioms they derive the maximum entropy principle for the selection of a default measure from a set of possible belief measures – without the taking into account of some given prior measure. The result of Paris and Vencovská is more general than that of Shore and Johnson in that they do not assume at the outset that the default selection is performed by minimizing some differentiable function (on the other hand, the scope of the Shore and Johnson result is wider than that of Paris and Vencovská because the latter one only refers to measures on finite spaces).

Not being concerned with updating a prior estimate of a distribution, the axioms of Paris and Vencovská cannot be brought into correspondence with rules for the logical behaviour of $\mid\!\approx^{\mathbf{R}}$ as readily as the Shore and Johnson axioms. Hence, one would rather choose the latter work as a basis for an axiomatic derivation of the minimum cross-entropy principle for default reasoning about probabilities.

Such a derivation of the minimum cross-entropy principle is of a completely different nature than the derivation by epistemic and statistical arguments given in sections 3.3 and 3.4. An argument based on an axiomatization of the default entailment relation can only be used to show that cross-entropy minimization is the adequate formalism for default reasoning about probabilities when it is taken for granted that at least one such formal process exists – an assumption that in itself is not corroborated by an axiomatic derivation. It might very well be that there are other axioms that are intuitively reasonable for default reasoning about probabilities, but are not satisfied by the minimum cross-entropy principle. In that case we would have to conclude that no completely adequate formal process exists. Deriving the minimum cross-entropy principle from a semantic model for the process of default reasoning about probabilities, on the other hand, provides valuable evidence that it does, in fact, not have counterintuitive logical properties, since these would have to correspond to flaws in the semantic model.

5.6 Axiomatizing Default Models

We show that the class of default models of $\phi \in \mathrm{L}^\beta$ is axiomatizable: there exists a theory $\mathrm{MinCE}(\phi) \subseteq \mathrm{L}^\beta$ in the vocabulary S of ϕ such that $(\mathfrak{M}, \nu_{\mathbf{e}}) \mid\!\approx \phi$ iff $(\mathfrak{M}, \nu_{\mathbf{e}}) \models_\beta \{\phi\} \cup \mathrm{MinCE}(\phi)$.

Formally, this is a very similar technique as used in logical default reasoning by circumscription (e.g. [McCarthy, 1980],[Lifschitz, 1986]). This approach to default reasoning is based on the concept of P-minimal models, i.e. models in which a specific predicate P expressing the property of being abnormal in a certain way has the smallest possible extension. A P-minimal model thereby captures the intuition that by default objects are assumed not to be abnormal.

The class of P-minimal models can be axiomatized by a second-order formula

$$\phi \wedge \forall X(\phi[\mathsf{P}/X] \to X \supseteq \mathsf{P}) \tag{5.64}$$

with X a predicate variable of the arity of P, and $\phi[\mathsf{P}/X]$ the formula obtained by replacing the predicate symbol P by the variable X. A first-order axiomatization of P-minimal models usually is not possible. Replacing the second-order formula (5.64) by a first-order schema

$$\phi \wedge (\phi[\mathsf{P}/\Psi] \to \Psi \supseteq \mathsf{P}), \tag{5.65}$$

where Ψ ranges over first-order formulas, only yields an axiomatization of models that are minimal with respect to first-order definable extensions of P.

The class of default models of $\phi \in \mathrm{L}^\beta$ is characterized by a formula whose outward appearance has much in common with (5.64):

$$\phi \wedge \forall \nu \in \Delta_{\mathrm{F}}\mathfrak{A}_{\mathbf{e}}(\phi[\mathrm{prob}/\nu] \rightarrow CE(\nu, \mu_{\mathbf{e}}) \geq CE(\mathrm{prob}, \mu_{\mathbf{e}})), \tag{5.66}$$

where $\phi[\mathrm{prob}/\nu]$ is the formula obtained by replacing occurrences of $\mathrm{prob}(\psi[\mathbf{e}])$ by $\nu(\psi[\mathbf{e}])$. Since $CE(\nu, \mu_{\mathbf{e}})$ is not defined when $\nu \not\ll \mu_{\mathbf{e}}$, and $(\mathfrak{M},\nu_{\mathbf{e}})$ is not a default model when $\nu_{\mathbf{e}} \not\ll \mu_{\mathbf{e}}$, we actually also have to add conditions of absolute continuity explicitly in (5.66):

$$\phi \wedge \mathrm{prob} \ll \mu_{\mathbf{e}} \wedge \forall \nu \in \Delta_{\mathrm{F}}\mathfrak{A}_{\mathbf{e}}((\nu \ll \mu_{\mathbf{e}} \wedge \phi[\mathrm{prob}/\nu]) \rightarrow CE(\nu, \mu_{\mathbf{e}}) \geq CE(\mathrm{prob}, \mu_{\mathbf{e}})). \tag{5.67}$$

The formula (5.67), however, is not yet in the syntax of L^β or predicate logic, and may only serve as an intuitive mould for the construction of such a formula. As it turns out, a sentence schema can be defined as an adequate formal counterpart of (5.67).

The reduction of (5.67) to an L^β- sentence schema is based on the fact that the quantifier "$\forall \nu \in \Delta_{\mathrm{F}}\mathfrak{A}_{\mathbf{e}}$" can be replaced by an infinite collection of quantifications "$\forall \nu \in \Delta_{\mathrm{F}}\mathfrak{A}$" with $\mathfrak{A}$ a finite algebra. Also, the condition $\mathrm{prob} \ll \mu_{\mathbf{e}}$ can be replaced by the conditions $\mathrm{prob}\restriction\mathfrak{A} \ll \mu_{\mathbf{e}}\restriction\mathfrak{A}$ with $\mathfrak{A}$ ranging over finite subalgebras of $\mathfrak{A}_{\mathbf{e}}$.

As in section 5.4, assume that ϕ contains n distinct subjective probability terms

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

For a belief structure $(\mathfrak{M},\nu_{\mathbf{e}})$ let $\mathfrak{A}^\chi$ be the subalgebra of $\mathfrak{A}_{\mathbf{e}}$ generated by the extensions of the ψ_i and one further formula $\chi\langle \boldsymbol{v}\rangle$.

Recall (cf. p. 114) that $\Pi_{\Delta_{\mathrm{F}}(\phi,\mathfrak{M})}(\mu_{\mathbf{e}})$ then consists just of those $\nu_{\mathbf{e}} \in \Delta_{\mathrm{F}}(\phi, \mathfrak{M})$ that satisfy (5.8) for all $\chi\langle \boldsymbol{v}\rangle \in \mathrm{L}^\sigma$. Also, $\nu_{\mathbf{e}}\restriction\mathfrak{A}^\chi \ll \mu_{\mathbf{e}}\restriction\mathfrak{A}^\chi$ for all χ clearly implies $\nu_{\mathbf{e}} \ll \mu_{\mathbf{e}}$. Hence, the conjuncts following ϕ in (5.67) may be replaced by the schema

$$\begin{aligned} &\mathrm{prob}\restriction\mathfrak{A}^\chi \ll \mu_{\mathbf{e}}\restriction\mathfrak{A}^\chi \wedge \forall \nu \in \Delta_{\mathrm{F}}\mathfrak{A}^\chi((\nu \ll \mu_{\mathbf{e}}\restriction\mathfrak{A}^\chi \wedge \phi[\mathrm{prob}/\nu]) \rightarrow \\ &\qquad\qquad CE(\nu, \mu_{\mathbf{e}}\restriction\mathfrak{A}^\chi) \geq CE(\mathrm{prob}\restriction\mathfrak{A}^\chi, \mu_{\mathbf{e}}\restriction\mathfrak{A}^\chi)) \quad (\chi\langle \boldsymbol{v}\rangle \in \mathrm{L}^\sigma). \end{aligned} \tag{5.68}$$

To represent instantiations of (5.68) in L^β, let

$$\Gamma := \left\{ \bigwedge_{i=1}^{n} \tilde{\psi}_i(\boldsymbol{v}) \wedge \tilde{\chi}\langle \boldsymbol{v}\rangle \mid \tilde{\psi}_i(\boldsymbol{v}) \in \{\psi_i(\boldsymbol{v}), \neg\psi_i(\boldsymbol{v})\}, \tilde{\chi}\langle \boldsymbol{v}\rangle \in \{\chi\langle \boldsymbol{v}\rangle, \neg\chi\langle \boldsymbol{v}\rangle\} \right\}. \tag{5.69}$$

Γ thus is a set of $N := 2^n$ expressions $\alpha_1(\boldsymbol{v}), \ldots, \alpha_N(\boldsymbol{v})$ defining the atoms of $\mathfrak{A}^\chi$ (or the empty set). It is assumed that we have some fixed schema according to which, given some $\phi \in \mathrm{L}^\beta$ with subjective probability terms $\mathrm{prob}(\psi_i[\mathbf{e}])$ and $\chi\langle \boldsymbol{v}\rangle \in \mathrm{L}^\sigma$, the elements $\alpha_j(\boldsymbol{v})$ of Γ are numbered, so that there is a uniform and effective way to compute for a given term $\mathrm{prob}(\psi_i[\mathbf{e}])$ in ϕ the set of indices j for which $\psi_i(\boldsymbol{v})$ occurs unnegated in $\alpha_j(\boldsymbol{v})$, i.e. the set

$$I(\phi, \psi_i) := \{j \in \{1, \ldots, N\} \mid \alpha_j(\boldsymbol{v}) = \psi_i(\boldsymbol{v}) \wedge \bigwedge_{\substack{k=1\\k\neq i}}^{n} \tilde{\psi}_k(\boldsymbol{v}) \wedge \tilde{\chi}\langle \boldsymbol{v}\rangle\}.$$

We can now formulate the L^β-building blocks for the formalization of (5.68). Let $\boldsymbol{x} = (x_1, \dots, x_N)$ be a tuple of field variables. Beginning with a translation for the continuity conditions, let

$$\zeta_{\phi,\chi}(\boldsymbol{x}) :\equiv \bigwedge_{j=1}^{N} ([\alpha_j(\boldsymbol{v})]_{\boldsymbol{v}} = 0 \to x_j = 0).$$

Next, the relativized quantification "$\forall \nu \in \Delta_{\mathrm{F}} \mathfrak{A}^{\chi}$" must be encoded by a formula expressing the property of $\boldsymbol{x}$ to be a probability measure on $\mathfrak{A}^{\chi}$:

$$\delta_{\phi,\chi}(\boldsymbol{x}) :\equiv \bigwedge_{j=1}^{N} x_j \geq 0 \wedge \sum_{j=1}^{N} x_j = 1 \wedge \bigwedge_{j=1}^{N} (\neg \exists \boldsymbol{v} \alpha_j(\boldsymbol{v}) \to x_j = 0).$$

The formula $CE(\nu, \mu_{\mathbf{e}} \restriction \mathfrak{A}^{\chi}) \geq CE(\mathrm{prob} \restriction \mathfrak{A}^{\chi}, \mu_{\mathbf{e}} \restriction \mathfrak{A}^{\chi})$ in proper syntax reads

$$\kappa_{\phi,\chi}(\boldsymbol{x}) :\equiv \forall y_1 \dots y_N \Bigg((\bigwedge_{j=1}^{N} ([\alpha_j(\boldsymbol{v})]_{\boldsymbol{v}} = 0 \vee y_j \cdot [\alpha_j(\boldsymbol{v})]_{\boldsymbol{v}} = 1)) \to \sum_{j=1}^{N} x_j \cdot \mathsf{ln}(x_j \cdot y_j) \geq \sum_{j=1}^{N} \mathrm{prob}(\alpha_j[\mathbf{e}]) \cdot \mathsf{ln}(\mathrm{prob}(\alpha_j[\mathbf{e}]) \cdot y_j) \Bigg).$$

In this formula $\kappa_{\phi,\chi}$, for the first time, we make use of having the symbol ln in our language. It has been introduced into L^β specifically for this use. With

$$X_i :\equiv \sum_{j \in I(\phi, \psi_i)} x_j \qquad (i = 1, \dots, n)$$

we can now encode the instance of (5.68) defined by $\chi\langle \boldsymbol{v} \rangle$ as the L^β-sentence

$$\begin{aligned} &\zeta_{\phi,\chi}(\mathrm{prob}(\alpha_1[\mathbf{e}]), \dots, \mathrm{prob}(\alpha_N[\mathbf{e}])) \wedge \\ &\quad \forall \boldsymbol{x}((\delta_{\phi,\chi}(\boldsymbol{x}) \wedge \zeta_{\phi,\chi}(\boldsymbol{x}) \wedge \phi[\mathrm{prob}(\psi_1[\mathbf{e}])/X_1, \dots, \mathrm{prob}(\psi_n[\mathbf{e}])/X_n]) \to \kappa_{\phi,\chi}(\boldsymbol{x})). \end{aligned}$$

Let $\mathrm{MinCE}(\phi)$ consist of all instantiations of this schema by formulas $\chi\langle \boldsymbol{v} \rangle \in \mathrm{L}^\sigma$.

By our assumption of a fixed algorithm that for every pair ϕ and χ produces the list $\alpha_1, \dots, \alpha_N$ of atoms of $\mathfrak{A}^{\chi}$ and the corresponding index sets $I(\phi, \psi_i)$, there exists an algorithm that for any given ϕ enumerates all the instantiations of $\mathrm{MinCE}(\phi)$.

Theorem 5.6.1 Let $\phi \in \mathrm{L}^\beta_{\mathrm{S},\mathbf{e}}$, $(\mathfrak{M}, \nu_{\mathbf{e}})$ a belief S-structure for $\mathbf{e}$. Then

$$(\mathfrak{M}, \nu_{\mathbf{e}}) \mathrel{|\!\!\approx} \phi \quad \text{iff} \quad (\mathfrak{M}, \nu_{\mathbf{e}}) \models_\beta \{\phi\} \cup \mathrm{MinCE}(\phi). \tag{5.70}$$

Proof: The proof really is rather obvious, because $\mathrm{MinCE}(\phi)$ is a straightforward encoding of the condition $\nu_{\mathbf{e}} \in \Pi_{\Delta_{\mathrm{F}}(\phi, \mathfrak{M})}(\mu_{\mathbf{e}})$. For completeness' sake we still go through the details of the proof, starting with the left to right direction.

Let $(\mathfrak{M}, \nu_{\mathbf{e}}) \mathrel{|\!\!\approx} \phi$. By definition $(\mathfrak{M}, \nu_{\mathbf{e}}) \models_\beta \phi$. Now let $\chi\langle \boldsymbol{v} \rangle \in \mathrm{L}^\sigma$ be given. It must be shown that $(\mathfrak{M}, \nu_{\mathbf{e}})$ is a feasible model of the instantiation of $\mathrm{MinCE}(\phi)$ by χ.

In the sequel we denote the tuple $(\text{prob}(\alpha_1[\mathbf{e}]), \dots, \text{prob}(\alpha_N[\mathbf{e}]))$ by $\mathbf{prob}$, and $(X_1, \dots, X_N)$ by $\boldsymbol{X}$. From $\nu_{\mathbf{e}} \ll \mu_{\mathbf{e}}$ it is immediate that

$$(\mathfrak{M}, \nu_{\mathbf{e}}) \models_\beta \zeta_{\phi,\chi}(\mathbf{prob}).$$

Now assume that

$$(\mathfrak{M}, \nu_{\mathbf{e}}) \models_\beta \exists \boldsymbol{x} \neg((\delta_{\phi,\chi}(\boldsymbol{x}) \wedge \zeta_{\phi,\chi}(\boldsymbol{x}) \wedge \phi[\mathbf{prob}/\boldsymbol{X}]) \to \kappa_{\phi,\chi}(\boldsymbol{x})),$$

i.e. there exists $r_1, \dots, r_N \in \mathbf{F}$ such that

$$((\mathfrak{M}, \nu_{\mathbf{e}}), \boldsymbol{x}/\boldsymbol{r}) \models_\beta \delta_{\phi,\chi}(\boldsymbol{x}) \wedge \zeta_{\phi,\chi}(\boldsymbol{x}) \wedge \phi[\mathbf{prob}/\boldsymbol{X}]) \wedge \neg\kappa_{\phi,\chi}(\boldsymbol{x}).$$

From the first two conjuncts it follows that

$$\mathfrak{M}(\alpha_j(\boldsymbol{v})) \mapsto r_j$$

induces a probability measure ν on $\mathfrak{A}^\chi$ with $\nu \ll \mu_{\mathbf{e}} \restriction \mathfrak{A}^\chi$. From

$$((\mathfrak{M}, \nu_{\mathbf{e}}), \boldsymbol{x}/\boldsymbol{r}) \models_\beta \phi[\mathbf{prob}/\boldsymbol{X}]$$

it follows that $\nu \in \Delta_{\mathrm{F}}(\phi, \mathfrak{M}) \restriction \mathfrak{A}^\chi$. Finally, for suitable $s_1, \dots, s_N \in \mathbf{F}$, we have

$$\begin{aligned} ((\mathfrak{M}, \nu_{\mathbf{e}}), \boldsymbol{x}/\boldsymbol{r}, \boldsymbol{y}/\boldsymbol{s}) \models_\beta \quad & \bigwedge_{j=1}^{N} ([\alpha_j(\boldsymbol{v})]_{\boldsymbol{v}} = 0 \vee y_j \cdot [\alpha_j(\boldsymbol{v})]_{\boldsymbol{v}} = 1)) \wedge \\ & \sum_{j=1}^{N} x_j \cdot \mathsf{ln}(x_j \cdot y_j) < \sum_{j=1}^{N} \text{prob}(\alpha_j[\mathbf{e}]) \cdot \mathsf{ln}(\text{prob}(\alpha_j[\mathbf{e}]) \cdot y_j)). \end{aligned}$$

Consider the two sums in this expression. Since we have both

$$(\mathfrak{M}, \nu_{\mathbf{e}}) \models_\beta \zeta_{\phi,\chi}(\mathbf{prob}) \quad \text{and} \quad (\mathfrak{M}, \boldsymbol{x}/\boldsymbol{r}) \models_\beta \zeta_{\phi,\chi}(\boldsymbol{x}),$$

for all indices j with $\mathfrak{M} \models [\alpha_j(\boldsymbol{v})]_{\boldsymbol{v}} = 0$ we get $(\mathfrak{M}, \nu_{\mathbf{e}})(\text{prob}(\alpha_j[\mathbf{e}])) = 0$ and $r_j = 0$. Hence, for all such indices

$$((\mathfrak{M}, \nu_{\mathbf{e}}), \boldsymbol{x}/\boldsymbol{r}, \boldsymbol{y}/\boldsymbol{s})(x_j \cdot \mathsf{ln}(x_j \cdot y_j)) = 0 \cdot \mathsf{ln}(0) = 0,$$
$$((\mathfrak{M}, \nu_{\mathbf{e}}), \boldsymbol{x}/\boldsymbol{r}, \boldsymbol{y}/\boldsymbol{s})(\text{prob}(\alpha_j[\mathrm{e}]) \cdot \mathsf{ln}(\text{prob}(\alpha_j[\mathrm{e}]) \cdot y_j)) = 0 \cdot \mathsf{ln}(0) = 0.$$

For all other $j \in \{1, \dots, N\}$, s_j is $1/\mathfrak{M}([\alpha_j(\boldsymbol{v})]_{\boldsymbol{v}})$, hence

$$\begin{aligned} ((\mathfrak{M}, \nu_{\mathbf{e}}), \boldsymbol{x}/\boldsymbol{r}, \boldsymbol{y}/\boldsymbol{s})(\sum_{j=1}^{N} x_j \cdot \mathsf{ln}(x_j \cdot y_j)) \quad &= \quad CE^{\mathrm{F}}(\nu, \mu_{\mathbf{e}} \restriction \mathfrak{A}^\chi), \\ ((\mathfrak{M}, \nu_{\mathbf{e}}), \boldsymbol{x}/\boldsymbol{r}, \boldsymbol{y}/\boldsymbol{s})(\sum_{j=1}^{N} \text{prob}(\alpha_j[\mathbf{e}]) \cdot \mathsf{ln}(\text{prob}(\alpha_j[\mathbf{e}]) \cdot y_j)) \quad &= \quad CE^{\mathrm{F}}(\nu_{\mathbf{e}} \restriction \mathfrak{A}^\chi, \mu_{\mathbf{e}} \restriction \mathfrak{A}^\chi), \end{aligned}$$

and therefore $CE^{\mathrm{F}}(\nu, \mu_{\mathbf{e}} \restriction \mathfrak{A}^\chi) < CE^{\mathrm{F}}(\nu_{\mathbf{e}} \restriction \mathfrak{A}^\chi, \mu_{\mathbf{e}} \restriction \mathfrak{A}^\chi)$. As $\Delta_{\mathrm{F}}(\phi, \mathfrak{M})$ is defined by constraints on $\mathfrak{A}^\chi$, it follows that $\nu_{\mathbf{e}} \notin \Pi_{\Delta_{\mathrm{F}}(\phi,\mathfrak{M})}(\mu_{\mathbf{e}})$, a contradiction. Since $\chi\langle \boldsymbol{v} \rangle$ has been arbitrary, we conclude that $(\mathfrak{M}, \nu_{\mathbf{e}}) \models_\beta \text{MinCE}(\phi)$.

The converse direction is similar: from $(\mathfrak{M}, \nu_{\mathbf{e}}) \not\approx \phi$ it follows that either $(\mathfrak{M}, \nu_{\mathbf{e}}) \not\models_\beta \phi$, in which case the right side of (5.70) is immediately seen to be false, or $(\mathfrak{M}, \nu_{\mathbf{e}}) \models_\beta \phi$ and $\nu_{\mathbf{e}} \notin \Pi_{\Delta_{\mathrm{F}}(\phi,\mathfrak{M})}(\mu_{\mathbf{e}})$. In the second case we either have $\nu_{\mathbf{e}} \not\ll \mu_{\mathbf{e}}$, yielding $(\mathfrak{M}, \nu_{\mathbf{e}}) \not\models \zeta_{\phi,\chi}(\mathbf{prob})$ for some χ, or $\nu_{\mathbf{e}} \ll \mu_{\mathbf{e}}$ and there exists $\nu^* \in \Delta_{\mathrm{F}}(\phi, \mathfrak{M})$, $\nu^* \ll \mu_{\mathbf{e}}$, with $CE^{\mathrm{F}}(\nu^*, \mu_{\mathbf{e}}) < CE^{\mathrm{F}}(\nu_{\mathbf{e}}, \mu_{\mathbf{e}})$. In the last case there exists some $\chi\langle \boldsymbol{v} \rangle \in \mathrm{L}^\sigma$ such that $CE^{\mathrm{F}}(\nu^* \upharpoonright \mathfrak{A}^\chi, \mu_{\mathbf{e}} \upharpoonright \mathfrak{A}^\chi) < CE^{\mathrm{F}}(\nu_{\mathbf{e}} \upharpoonright \mathfrak{A}^\chi, \mu_{\mathbf{e}} \upharpoonright \mathfrak{A}^\chi)$. The variable assignment

$$\gamma(x_j) := \nu^*(\mathfrak{M}(\alpha_j(\boldsymbol{v})) \qquad (j = 1, \ldots, N)$$

then falsifies

$$(\delta_{\phi,\chi}(\boldsymbol{x}) \wedge \zeta_{\phi,\chi}(\boldsymbol{x}) \wedge \phi[\mathbf{prob}/\boldsymbol{X}]) \to \kappa_{\phi,\chi}(\boldsymbol{x}).$$

□

With theorem 5.6.1 a completeness result for $\mathrel{|\!\!\approx}$ now follows: From (5.70) and theorem 5.3.1 we get for $\phi, \psi \in \mathrm{L}^\beta$:

$$\phi \mathrel{|\!\!\approx} \psi \quad \text{iff} \quad \{\phi^*\} \cup \mathrm{MinCE}(\phi)^* \cup \mathrm{AX}(\mathbf{e}) \models \psi^*.$$

Thus, the default inference relation $\mathrel{|\!\!\approx}$ is captured completely within a first-order formalism. A complete proof system for $\mathrel{|\!\!\approx}$ therefore is given by the algorithm that for every $\phi \in \mathrm{L}^\beta$ enumerates all instantiations of $\mathrm{MinCE}(\phi)$, the translation rules for $(\cdot)^*$ and $(\cdot)^{-1}$, and a proof system for first-order logic. As was the case for the reduction of $\models_\sigma$ and $\models_\beta$ to first-order inferences, the translations $(\cdot)^*$ and $(\cdot)^{-1}$ can be eliminated from this proof system by extending a first-order system by substitution rules for probability terms, adding the axiom system $\mathrm{AX}(\mathbf{e})^{-1}$, and then work directly within the language L^β.

Chapter 6

Comparisons

6.1 The Work of Bacchus et al.

Under this heading three distinct formalisms have to be distinguished: Bacchus's logic LP, the logic $\mathscr{L}_3^=$ due to Bacchus and Halpern, and the random worlds formalism by Bacchus, Grove, Halpern, and Koller.

LP is the logic for statistical probabilities described in [Bacchus, 1990a] that has already been discussed in section 2.3, and is essentially the same as $\mathscr{L}^\sigma$. The more substantial differences between previous formalisms and the one introduced in the present work lie in the handling of subjective probabilities.

6.1.1 The Logic $\mathscr{L}_3^=$

The language $\mathscr{L}_3^=$ ([Bacchus, 1990b],[Halpern, 1990]) has already been mentioned in section 5.1. It differs from the language L^β by not having the special function symbol ln as a fixed part of the language, and, more importantly, by not basing the formation of subjective probability terms on a special set of event symbols $\mathbf{e}$. Rather, for every formula $\phi(\boldsymbol{v}) \in \mathscr{L}_3^=$,

$$\mathrm{prob}(\phi(\boldsymbol{v}))$$

is a subjective probability term in $\mathscr{L}_3^=$. Specifically, ϕ is allowed to contain any type of nestings of the subjective probability operator prob() and the statistical quantifier $[\cdot]$. The resulting term $\mathrm{prob}(\phi(\boldsymbol{v}))$ depends on the free variables of ϕ.

One example for a representation that makes use of the more flexible syntax of $\mathscr{L}_3^=$ compared to L^β already has been given in section 5.1:

$$[\mathrm{prob}(\mathsf{Betterf}v) \geq 0.9]_v \leq 0.1.$$

The language $\mathscr{L}_3^=$ is interpreted by combining the notion of first-order structures with possible worlds semantics. A "type-3 probability structure" for the interpretation of $\mathscr{L}_3^=$ then has the form

$$\mathfrak{M} = (D, S, \pi, \mu_D, \mu_S)$$

with D a domain, S a set of possible worlds (each with domain D), π a mapping that assigns to every $s \in S$ an interpretation of the symbols in the given vocabulary, μ_D a probability measure on D, and μ_S a probability measure on S. The two probability measures are assumed to be real-discrete, so that the issue of measurability does not arise, and the one-dimensional measure μ_D can be extended to n dimensions in the canonical way by the product measure μ_D^n. Formulas $\phi(\boldsymbol{v}) \in \mathscr{L}_3^=$ then are given a truth value in $\mathfrak{M}$ at a specific world $s \in S$ with respect to a variable assignment γ (which is independent of s) by a straightforward inductive definition. The key elements of this definition are the interpretations of the two types of probability terms, given by

$$(\mathfrak{M}, s, \gamma)([\phi(\boldsymbol{v},\boldsymbol{w},\boldsymbol{x})]_{\boldsymbol{w}}) = \mu_D^{|\boldsymbol{w}|}(\{\boldsymbol{d} \in D^{|\boldsymbol{w}|} \mid (\mathfrak{M}, s, \gamma[\boldsymbol{w}/\boldsymbol{d}]) \models \phi\}),$$
$$(\mathfrak{M}, s, \gamma)(\mathrm{prob}(\phi(\boldsymbol{v},\boldsymbol{x}))) = \mu_S(\{s' \in S \mid (\mathfrak{M}, s', \gamma) \models \phi(\boldsymbol{v},\boldsymbol{x})\}).$$

By this definition, only the interpretation of statistical probability terms depends on the current world. Subjective probability terms, on the other hand, have the same interpretation in every world s.

The main source from which the ease springs with which complex combinations of probabilistic statements can be interpreted in this framework is the restriction to real-discrete probabilities, which makes every subset of the domain and every set of possible worlds measurable. This relieves us of the need to always take care for there to be an adequate algebraic structure on which probabilities are defined.

If for L^β we only had been interested in defining the semantics of feasible models, i.e. the logic $\mathscr{L}^\beta$, then we might also have used the unrestricted syntax of $\mathscr{L}_3^=$. Assuming an extended closure condition for belief structures (which becomes trivially satisfied if the restriction to real-discrete structures is being made), then generalized probability terms $\mathrm{prob}(\phi(\boldsymbol{v},\boldsymbol{w},\boldsymbol{x})[\boldsymbol{v}/\mathbf{e}])$ that may contain free variables and nested occurrences of prob() also define measurable subsets of $M^{|\mathbf{e}|}$, and can thus be interpreted by a measure $\nu_{\mathbf{e}}$ on $M^{|\mathbf{e}|}$.

Only for the definition of default models do the restrictions of the syntax of L^β become really essential, because they ensure that $\Delta_{\mathrm{F}}(\phi, \mathfrak{M})$ is defined by constraints on a finite algebra.

On the level of strict logical inference, random event semantics therefore is not inherently different from possible worlds semantics with respect to the scope of probabilistic statements that meaning can be given to. The two semantic concepts, however, lead to a somewhat different behaviour of the logics that use them as their basis.

For want of a better name (Halpern does not supply one), in the sequel we extend the denotation $\mathscr{L}_3^=$ also to the logic defined by the language $\mathscr{L}_3^=$ and its possible worlds semantics.

As a first example illustrating the distinct logical properties of $\mathscr{L}^\beta$ and $\mathscr{L}_3^=$, consider the formula

$$\mathrm{prob}(\forall v(\mathsf{P}\, v \rightarrow \mathsf{Q}\, v)) = r, \tag{6.1}$$

which is both in L^σ and in $\mathscr{L}_3^=$. In $\mathscr{L}_3^=$ the probability term $\mathrm{prob}(\forall v(\mathsf{P}\, v \rightarrow \mathsf{Q}\, v))$ designates the probability of the set of worlds in which $\forall v(\mathsf{P}\, v \rightarrow \mathsf{Q}\, v)$ is true. In suitable type-3 structures this probability can have any value, so that in $\mathscr{L}_3^=$ (6.1) is satisfiable for every $r \in [0,1]$. In $\mathscr{L}^\beta$, on the other hand, $\mathrm{prob}(\forall v(\mathsf{P}\, v \rightarrow \mathsf{Q}\, v))$ is interpreted either as 0 or as 1: there not occurring any event symbols in $\forall v(\mathsf{P}\, v \rightarrow \mathsf{Q}\, v)$, this formula in $M^{|\mathbf{e}|}$ defines either the empty set or the full set $M^{|\mathbf{e}|}$ – depending on whether $\forall v(\mathsf{P}\, v \rightarrow \mathsf{Q}\, v)$ holds in (M, I) or not.

Generally, a formula like $\forall v(\mathsf{P}\, v \rightarrow \mathsf{Q}\, v)$ not containing any event symbols can only have probability 0 or 1 in $\mathscr{L}^\beta$, but (unless it is a tautology or unsatisfiable) any probability in $\mathscr{L}_3^=$.

(6.1) illustrates the possible worlds semantics' greater versatility, compared to random event semantics, in assigning nontrivial subjective probabilities to formulas of any syntactic form. A second respect in which possible worlds semantics is more flexible is the independence from any given (definite or statistical) domain information with which subjective probabilities can be assigned. An example for how this can be put to use is

$$\begin{gathered}\exists^{=1} v(\mathsf{ActorA}\, v \wedge \mathsf{ActorB}\, v) \quad [\mathsf{A}\, v \mid \mathsf{ActorA}\, v]_v = 0.8 \quad [\mathsf{A}\, v \mid \mathsf{ActorB}\, v]_v = 0.2 \\ \mathrm{prob}(\mathsf{ActorA}\, \mathsf{f} \wedge \mathsf{ActorB}\, \mathsf{f}) = 1.\end{gathered} \tag{6.2}$$

where $\exists^{=1}$ abbreviates "there exists exactly one". Let ϕ be the conjunction of the formulas (6.2) and the formula $\mathrm{prob}(\mathsf{A}\, \mathsf{f}) = 0.5$. ϕ formalizes the example given at the end of section 3.2 (p. 69). It is satisfiable in $\mathscr{L}_3^=$, but not in $\mathscr{L}^\beta$: in any feasible model $\mathfrak{M}$ of (6.2) the set $\mathfrak{M}(\mathsf{ActorA}\, v \wedge \mathsf{ActorB}\, v)$ consists of just one element a, with $\nu_{\mathrm{f}}(\{a\}) = 1$. Depending on whether $a \in I(\mathsf{A})$, then either $\nu_{\mathrm{f}}(\mathsf{A}) = 1$ or $= 0$, so that $\mathrm{prob}(\mathsf{A}\, \mathsf{f}) = 0.5$ can not be true in $\mathfrak{M}$.

While sometimes it may be useful, as in this example, to express degrees of belief that are somewhat at odds with the given domain information, the semantics of $\mathscr{L}_3^=$ interprets degrees of belief in such a completely detached manner that even subjective probabilities become admissible that should be regarded as inconsistent with the given facts:

$$\neg\exists v \mathsf{P}\, v \;\wedge\; \mathrm{prob}(\mathsf{P}\, \mathsf{e}) = 1,$$

according to $\mathscr{L}_3^=$, is satisfiable at a world in a type-3 structure that has zero probability, and at which $\neg\exists v \mathsf{P}\, v$ holds, whereas $\mathsf{P}\, \mathsf{e}$ is true at all worlds with positive probability. This example highlights the fact that $\mathscr{L}_3^=$ does not support any inferences of degrees of belief from the given objective information, particularly default reasoning about probabilities.

Halpern [1990] also considers the particular class of type-3 structures in which predicate and function symbols are *rigid*, i.e. have the same interpretation in every possible world (which then differ only with respect to the interpretation of constant symbols).

With this restriction, the possible worlds semantics becomes equivalent to random event semantics (with every constant symbol e in $\mathscr{L}_3^=$considered an event symbol), because then formulas $\phi(\boldsymbol{v})$ not containing any constant symbols define the same set $\mathfrak{M}(\phi(\boldsymbol{v}))$ in every possible world of a type-3 structure $\mathfrak{M}$, so that via

$$\nu_{\mathbf{e}}(\mathfrak{M}(\phi(\boldsymbol{v}))) = \mu_S(\{s' \in S \mid (\mathfrak{M}, s', \gamma[\boldsymbol{v}/\mathbf{e}]) \models \phi(\boldsymbol{v})\})$$

a correspondence is defined between probability measures $\nu_{\mathbf{e}}$ on the algebra of these sets, and probability measures μ_S over possible worlds.

6.1.2 The Random Worlds Method

The random worlds method developed by Bacchus, Grove, Halpern, and Koller ([Bacchus *et al.*, 1992],[Grove *et al.*, 1992a],[Grove *et al.*, 1992b], [Bacchus *et al.*, 1993],[Bacchus *et al.*, 1994a]) is a formalism for deriving degrees of belief from statistical information. The general framework has been brought to bear on two separate issues: the main thrust of [Bacchus *et al.*, 1993] is to use the random worlds method as a probabilistic formalization of (logical) default reasoning, while in the other sources cited it is investigated primarily as a system for deriving degrees of belief within the whole scope [0,1] of values. This second application being much closer to what we have called default reasoning about probabilities, we will concentrate on that second perspective on the random worlds method.

In brief outline, the random worlds method works as follows. For some given vocabulary S and for every $N \geq 1$ let $\mathfrak{M}_N$ be the particular type-3 structure with the set of possible worlds being the set $\mathcal{W}_N$ of all possible S-structures over the domain $D = \{1, \ldots, N\}$ equipped with the uniform probability measure, i.e. $\mu_D(\{i\}) = 1/N$ for $i \in D$, and $\mu_{\mathcal{W}_N}$ the uniform distribution on $\mathcal{W}_N$, i.e. $\mu_{\mathcal{W}_N}(\{s\}) = 1/\mid \mathcal{W}_N \mid$ for all $s \in \mathcal{W}_N$. Possible S-structures here are defined by treating the elements of D as distinguishable individuals, so that when S contains a single constant symbol a, for instance, the structure in which a is interpreted by the domain element i is distinguished from the structure in which a is interpreted by the domain element $j \neq i$, and $\mathcal{W}_N$ contains exactly N possible worlds (even though they are all isomorphic).

For two sentences $\phi, \theta \in \mathscr{L}_3^=$ not containing the prob()- operator (equivalently: $\phi, \theta \in \mathrm{L}^\sigma$) define

$$\mathrm{Pr}_N(\phi \mid \theta) := \frac{\mathfrak{M}_N(\mathrm{prob}(\phi \wedge \theta))}{\mathfrak{M}_N(\mathrm{prob}(\theta))}$$

which, by the definition of $\mathfrak{M}_N$ is just the number of worlds in which $\phi \wedge \theta$ is true divided by the number of worlds in which θ is true (we here gloss over the problem that in order to avoid unwanted dependencies of the satisfiability of atomic field formulas on the precise domain size N, the conditions for the validity of an atomic field formula at a given world in $\mathcal{W}_N$ must be somewhat relaxed by only demanding an approximate satisfaction of the (in-) equality statement made by the formula. The details are given in [Grove *et al.*, 1992b].).

Arguing that the actual domain that is described by ϕ and θ has some unknown, large, but finite size, the random worlds method proposes the limit

$$\mathrm{Pr}_\infty := \mathit{lim}_{\,N \to \infty} \mathrm{Pr}_N(\phi \mid \theta) \tag{6.3}$$

(provided this limit exists) as the degree of belief in ϕ that should be derived from θ.

As a first example, consider

$$\begin{aligned} \theta &:\equiv [\mathsf{P}\, v \mid \mathsf{Q}\, v]_v \geq 0.8 \wedge \mathsf{Q}\,\mathsf{a} \\ \phi &:\equiv \mathsf{P}\,\mathsf{a}. \end{aligned}$$

As shown in [Grove *et al.*, 1992b], when S is a monadic vocabulary, for large N, the subset of $\mathcal{W}_N$ that satisfies a given set of statistical constraints gets dominated by those worlds where the statistical distribution on the algebra generated by the interpretation of the predicate symbols is close to the maximum entropy solution of the given constraints. For the given example this means that for large N eventually almost all worlds in $\mathcal{W}_N$ that satisfy θ will actually satisfy

$$\theta' :\equiv [\mathsf{P}\, v \mid \mathsf{Q}\, v]_v = 0.8 \wedge \mathsf{Q}\,\mathsf{a}$$

because 0.8 is the conditional probability of P given Q according to the maximum entropy distribution on the four-element algebra generated by the interpretations of P and Q under the constraint θ (again, note that θ' must be seen to be satisfied at a world of size N if the equation $[\mathsf{P}\,v \mid \mathsf{Q}\,v]_v = 0.8$ actually only is true within a small admissible tolerance). Hence, by the random worlds method we derive

$$\Pr_\infty(\theta' \mid \theta) = 1.$$

Also, since for any N the fraction of models of θ' in which $\mathsf{P\,a}$ holds is 0.8, we obtain

$$\Pr_\infty(\phi \mid \theta) = 0.8.$$

As explained so far, the random worlds method only allows to derive degrees of belief in certain statements on the basis of some given objective information. It thus does not address the main issue that motivated the development of $\mathscr{L}^\beta_{\text{def}}$, namely the combination of two types of probabilistic information. In [Bacchus *et al.*, 1994a] it has been shown, however, how the random worlds method can be extended to also use prior degrees of belief as input. Before we here discuss this extension, it is still useful to first explore some of the basic characteristics of the simple form of random worlds reasoning, all of which are shared by the refined version.

A substantial difference between the derivations of degrees of belief by the random worlds method and in $\mathscr{L}^\beta_{\text{def}}$ becomes clear from the simple example above: random worlds inherently makes default assumptions about statistical probabilities, which to a large extent is avoided in $\mathscr{L}^\beta_{\text{def}}$. Also, whenever degrees of belief are derived by the random worlds method, these are point-valued, while in $\mathscr{L}^\beta_{\text{def}}$ usually only interval-valued degrees of belief are found, their vagueness typically reflecting the incompleteness of the given statistical information.

Another relevant distinguishing feature between random worlds and $\mathscr{L}^\beta_{\text{def}}$ is the commitment of the former to finite domains. Clearly, this is a technical necessity, because on the set of all possible S-structures over an infinite domain there exists no uniform probability distribution which then might have been argued to be the most natural probability measure for assigning degrees of belief. Bacchus et al. also stipulate that this restriction to finite domains and the limiting process in (6.3) is epistemologically adequate because "In our context, we can assume that the 'true world' has a finite domain, say size N", and "Typically, we know [N not] exactly. All we know is that N is 'large' [...]" [Bacchus *et al.*, 1994b].

While it may be correct that the "true world" is finite, this does not necessarily justify the limitation of abstract models of the true world to finite domain models: an imperfect and somewhat idealized description of the real world that we are likely to find as a knowledge base may very well only possess infinite models, even though it actually is meant to describe a finite set of objects.

Consider the following self-explanatory example.

$$\forall uw((\mathsf{mother}(u) = w \vee \mathsf{father}(u) = w) \rightarrow \mathsf{Ancestor}\,wu) \tag{6.4}$$

$$\forall uvw((\mathsf{Ancestor}\,uv \wedge \mathsf{Ancestor}\,vw) \rightarrow \mathsf{Ancestor}\,uw) \tag{6.5}$$

$$\forall uv(\mathsf{Ancestor}\,uv \rightarrow \neg\mathsf{Ancestor}\,vu) \tag{6.6}$$

$$\forall u([\mathsf{Female}\,w \mid \mathsf{Ancestor}\,wu)]_w = 0.5) \tag{6.7}$$

$$\mathsf{Ancestor\,ba} \tag{6.8}$$

Let θ be the conjunction of (6.4)-(6.8). Clearly, θ will be seen as a perfectly valid description of certain relationships among human beings, even though the number of people that have ever lived is (large but) finite, yet θ only has infinite models. An intuitive inference one might want to derive from θ is that $\mathsf{Female}\,\mathsf{b}$ has a probability of 0.5. This result is easily obtained in $\mathscr{L}^{\beta}_{\mathrm{def}}$ when b (and, optionally, a as well) is interpreted as an event symbol. The random worlds method, on the other hand, can not generate any degrees of belief from θ.

Non-satisfiability over finite domains is only one problem that can prevent the random worlds method from producing the inferences one would like to expect. The complementary problem is that from a knowledge base θ that is satisfiable over all sufficiently large finite domains, the random worlds method will produce certain belief in every formula ϕ for which $\theta \rightarrow \phi$ is a tautology over finite domains. If θ', for example is defined as the conjunction of (6.4), (6.5), (6.7), and (6.8), then θ' does have finite models, and, in fact, we can derive

$$\mathrm{Pr}_{\infty}(\mathsf{Female}\,\mathsf{b} \mid \theta') = 0.5.$$

But also

$$\mathrm{Pr}_{\infty}(\exists uv(\mathsf{Ancestor}\,uv \wedge \mathsf{Ancestor}\,vu) \mid \theta') = 1,$$

which is not really what we would like to infer.

The extension of random worlds that also incorporates reasoning from prior subjective beliefs works with the full language $\mathscr{L}^{=}_{3}$. Let $\mathfrak{M}_N$ be defined as above and consider $\theta \in \mathscr{L}^{=}_{3}$. Usually, θ will not be satisfied in every world $s \in \mathcal{W}_N$, so that $\mathfrak{M}_N(\mathrm{prob}(\theta)) < 1$. For some other probability measure $\tilde{\mu}_{\mathcal{W}_N}$ on $\mathcal{W}_N$, and the corresponding type-3 structure $\tilde{\mathfrak{M}}_N$, defined just like $\mathfrak{M}_N$ with $\mu_{\mathcal{W}_N}$ replaced by $\tilde{\mu}_{\mathcal{W}_N}$, it can be true that $\tilde{\mathfrak{M}}_N(\mathrm{prob}(\theta)) = 1$. Using somewhat familiar notation, let

$$\Delta(\theta, N) := \{\tilde{\mu}_{\mathcal{W}_N} \in \Delta\mathcal{W}_N \mid \tilde{\mathfrak{M}}_N(\mathrm{prob}(\theta)) = 1\}.$$

If $\Delta(\theta, N)$ contains a unique element $\mu^{\theta}_{\mathcal{W}_N}$ that maximizes entropy within $\Delta(\theta, N)$ (and hence minimizes cross-entropy with respect to the uniform measure $\mu_{\mathcal{W}_N}$), for $\phi \in \mathscr{L}^{=}_{3}$ define

$$\mathrm{Pr}_N(\phi \mid \theta) := \mathfrak{M}^{\theta}_N(\mathrm{prob}(\phi)),$$

where $\mathfrak{M}^{\theta}_N$ is $\mathfrak{M}_N$ with $\mu_{\mathcal{W}_N}$ replaced by $\mu^{\theta}_{\mathcal{W}_N}$. Provided the limit exists, define

$$\mathrm{Pr}_{\infty}(\phi \mid \theta) := \mathit{lim}_{\,N\rightarrow\infty}\mathrm{Pr}_N(\phi \mid \theta).$$

The properties of this inference rule have not yet been explored at any detail. We will here examine a few examples that may shed some light on the behaviour of this generalized random worlds method, and how it compares to $\mathscr{L}^{\beta}_{\mathrm{def}}$.

The first example discusses a knowledge base that illustrates the main advantage of random worlds over $\mathscr{L}^{\beta}_{\mathrm{def}}$: the ability to assign nontrivial probabilities to propositions not featuring any constant (=event) symbols.

Example 6.1.1 Let

$$\theta :\equiv \mathrm{prob}(\forall v(\mathsf{P}\,v \rightarrow \mathsf{Q}\,v)) \geq 0.6 \wedge \mathsf{P}\,\mathsf{a}.$$

What can be inferred about the probability of $\mathsf{Q\,a}$? In $\mathscr{L}^{\beta}_{\mathrm{def}}$ the answer is quick and decisive: the objective statement $\forall v(\mathsf{P}\,v \rightarrow \mathsf{Q}\,v)$ in a belief structure can only be assigned probability 0 or 1 (cf. p. 136). The first possibility being ruled out by θ, we derive prob$(\forall v(\mathsf{P}\,v \rightarrow \mathsf{Q}\,v))$=1, and consequently prob$(\mathsf{Q\,a})$=1.

Now consider the type-3 structures $\mathfrak{M}_N$ for the vocabulary $\{\mathsf{P}, \mathsf{Q}, \mathsf{a}\}$. The proportion of worlds that satisfy $\forall v(\mathsf{P}\,v \rightarrow \mathsf{Q}\,v)$ becomes negligible as N grows large, so that the maximum entropy measure on $\mathcal{W}_N$ that is consistent with θ assigns the minimal possible probability 0.6 to worlds satisfying this sentence, and distributes the remaining probability 0.4 evenly among worlds in which $\mathsf{P\,a} \wedge \neg(\forall v(\mathsf{P}\,v \rightarrow \mathsf{Q}\,v))$ is true. This latter set of worlds becomes dominated by worlds in which approximately half the elements of P also belong to Q, so that here in approximately half the number of worlds $\mathsf{Q\,a}$ will be true. Hence

$$\mathrm{Pr}_{\infty}(\mathsf{Q\,a} \mid \theta) = 0.6 + 0.4 \cdot 0.5 = 0.8.$$

Next it is shown how random worlds handles example 5.4.18. As can be expected from a formalism that makes default assumptions about the statistical distribution, and is related to entropy maximization, it proves to be representation dependent.

Example 6.1.2 Let ϕ_1 be given as in example 5.4.18 by the conjunction of the sentences (5.43). The maximum entropy distribution on the algebra generated by the relation symbols A and B under the given constraints for the conditional probability of B, is computed to assign a probability ≈ 0.57 to the predicate A.

For large N, therefore, in the subset of $\mathcal{W}_N$ that satisfies the statistical constraints, approximately a proportion of 57% of worlds are a model of $\mathsf{A}\,\mathrm{e}$. Since this is consistent with the constraint prob$(\mathsf{A}\,\mathrm{e}) \geq 0.4$, the maximum entropy distribution on $\mathcal{W}_N$ that assigns probability 1 to ϕ_1 is given by the uniform distribution concentrated on the subset of $\mathcal{W}_N$ containing the worlds that satisfy the statistical constraints, and consequently

$$\mathrm{Pr}_{\infty}(\mathsf{A}\,\mathrm{e} \mid \phi_1) \approx 0.57.$$

Also, the statistical probability of B in the maximum entropy distribution on the algebra of A and B under the constraints is approximately 0.21, so that

$$\mathrm{Pr}_{\infty}(\mathsf{B}\,\mathrm{e} \mid \phi_1) \approx 0.21.$$

Now consider ϕ_2, the conjunction of the formulas (5.44). Here the maximum entropy distribution on the algebra generated by $\mathsf{A}_1, \mathsf{A}_2$, and B under the two statistical constraints on the conditional probability of B assigns a probability of 0.8 to the union of A_1 and A_2, and a probability of 0.26 to B. Hence

$$\mathrm{Pr}_{\infty}(\mathsf{A}_1\,\mathrm{e} \vee \mathsf{A}_2\,\mathrm{e} \mid \phi_2) = 0.8, \qquad \mathrm{Pr}_{\infty}(\mathsf{B}\,\mathrm{e} \mid \phi_2) = 0.26.$$

The last example we here present once more illustrates the consequences of the random worlds method's preference for models with particular statistical distributions.

Example 6.1.3 Let ϕ be defined as in example 5.4.14. It is readily verified that for large N the measure with maximal entropy on $\mathcal{W}_N$ that assigns probability to models of ϕ is given by assigning probability 1 to models of σ_1, and probability 0.3 to models of $\mathsf{R}_1\,\mathsf{t} \vee \mathsf{R}_2\,\mathsf{t} \vee \mathsf{R}_3\,\mathsf{t}$. Hence,

$$\mathrm{Pr}_\infty(\sigma_1 \mid \phi) = 1,$$

and also

$$\mathrm{Pr}_\infty(\mathsf{R}_1\,\mathsf{t} \mid \phi) = 0.1.$$

Hence, where in $\mathscr{L}^{\beta}_{\mathrm{def}}$ two different degrees of belief are derived, conditioned on the two different statistical hypotheses σ_1 and σ_2, the random worlds method decides that σ_1 is the better statistical hypothesis, and assigns degrees of belief accordingly.

A really comprehensive study of the relationship between possible worlds semantics and the random worlds method on the one hand, and random event semantics and the default inference relation $\models\!\!\approx$ on the other, is far beyond what here can be undertaken. As a preliminary conclusion we can summarize the main distinguishing features of the two approaches as follows: the great advantage of random worlds semantics is their ability to interpret (by non-trivial probability values) any kind of probabilistic first-order statement, without the need to associate a distinguished set of event symbols with subjective probabilities. In general, possible worlds semantics do not support a connection of objective statements (statistical or deterministic) with degrees of belief. The random worlds method is a technique to nonetheless achieve such a connection. There are two main sources for the differences between results obtained by the random worlds method and by $\mathscr{L}^{\beta}_{\mathrm{def}}$: first, by always implicitly performing a "completion" of partial statistical information, the random worlds method will always produce point-valued degrees of belief (if any), whereas in $\mathscr{L}^{\beta}_{\mathrm{def}}$ usually only interval-valued degrees of belief are derivable. The two systems thereby represent two different preferences with regard to the tradeoff between inferential strength on the one hand, and epistemological justifiability on the other. The second source for the qualitative difference of the two systems is the random worlds method's dependence on finite models, which must be seen as a serious limitation.

All these considerations address the semantical differences only. There also is a very substantial difference with respect to proof theory: while for $\mathscr{L}^{\beta}_{\mathrm{def}}$ there exists a complete proof system, for the random worlds method it is only known for some special cases how to actually derive degrees of belief (monadic languages: [Grove *et al.*, 1992b], knowledge bases of some specific structures: [Bacchus *et al.*, 1992]). The general inference problem for the random worlds method is incomplete ([Grove *et al.*, 1992a]). Using generalize field-valued probabilities here offers no solution, because (6.3) necessarily can only define a real number.

6.2 The Work of Paris and Vencovská

Paris and Vencovská ([1989],[1992]) propose a framework for probabilistic reasoning from knowledge bases that can be viewed as the propositional version of L^{β}: they consider statements that, adjusted to our notation, have the form

$$\sum_{i=1}^{k} a_i[\phi_i] \quad = \quad r \tag{6.9}$$

$$\mathrm{prob}(\psi(\mathrm{e})) \quad = \quad s. \tag{6.10}$$

where ϕ_i, ψ are propositional formulas over propositional variables $A_1, \ldots, A_n$, e is an event symbol, and $a_i, r, s \in \mathbf{R}$.

Paris and Vencovská propose an interpretation of these formulas that is diametrically opposed to the possible worlds semantics (and much closer in spirit to random event semantics) in that a very strong link is established between the two types (6.9) and (6.10) of probabilistic statements. In fact, (6.10) is seen as a special case of (6.9): a degree of belief that a specific event (object) e has property ψ, by Paris and Vencovská is interpreted as the statistical probability of ψ in an ideal reference class S_{e} of events "similar to" e. This interpretation of subjective probability, which we shall call *reference class semantics*, clearly is very much in line with the frequentistic point of view mentioned in section 3.2 (p. 68).

By translating a statement (6.10) into the form

$$[\psi \mid S_{\mathrm{e}}] = s, \tag{6.11}$$

the problem of default reasoning about probabilities is integrated into the general problem of probabilistic inferences from information about a single type of probabilities.

Since the set of events "similar to" e can be supposed to be very small, Paris and Vencovská also propose to add a constraint

$$[S_{\mathrm{e}}] = \epsilon \tag{6.12}$$

to the knowledge base, where $\epsilon > 0$ is assumed to be small. A knowledge base containing statements of the form (6.9) and (6.10) thus is finally transformed into a knowledge base containing the statements (6.9), the transformations (6.11) of (6.10), and the additional constraint (6.12).

In [1992] Paris and Vencovská consider different inference processes that can be applied to this knowledge base, most eminently the maximum entropy inference process. Particularly, they show that when $\mu_{\mathrm{me},\epsilon}$ is the maximum entropy solution for the constraints in the knowledge base (with ϵ as in (6.12)), and $\mu^{S_{\mathrm{e}}}_{\mathrm{me},\epsilon}$ is $\mu_{\mathrm{me},\epsilon}$ conditioned on S_{e}, then, in the limit for $\epsilon \to 0$, $\mu^{S_{\mathrm{e}}}_{\mathrm{me},\epsilon}$ is just the minimum cross-entropy solution of the constraints (6.10) with respect to the maximum entropy solution of the statistical constraints (6.9).

Since in [1990] Paris and Vencovská gave an axiomatic derivation of the maximum entropy inference principle (cf. p. 128), this is seen as a justification for the minimum cross-entropy principle (for default reasoning about probabilities).

It has already been remarked in section 5.5 that a derivation of the minimum cross-entropy principle of this kind has a distinctly different quality from the one presented in chapter 3.

As compared to Shore and Johnson's [1980] derivation of the minimum cross-entropy principle, the justification given by Paris and Vencovská is somewhat flawed, because it depends on the application of the maximum entropy principle to statistical information. It here has been argued before (cf. p. 7) that it is much harder to justify the application of an inference process to (objective) statistical probabilities than it is to justify such an application to subjective probabilities. Indeed, Paris and Vencovská seem to share this point of view to some extent, because in [1990] they expressly state that the axiomatization of an inference process they present is meant to deal with "some kind of subjective probabilities or degrees of belief", not with statistical probability measures. Thus, in [Paris and Vencovská, 1990] the maximum entropy principle has been derived for an intended scope of application not covering the use made of it in [Paris and Vencovská, 1992].

The use of reference class semantics certainly is not limited to a propositional context. In fact, it is immediately clear, how to extend this semantics for interpreting the language L^{β}: subjective probability terms $\mathrm{prob}(\psi(\mathbf{e}))$ ($\psi \in \mathrm{L}^{\sigma}$) here, too, can be understood as conditional statistical probability terms of the form $[\psi(\boldsymbol{v}) \mid S_{\mathbf{e}}\,\boldsymbol{v}]_{\boldsymbol{v}}$, with $S_{\mathbf{e}}$ the reference class for $\mathbf{e}$.

Such reference class semantics for L^{β} would establish an even greater link between domain information and degrees of belief than random event semantics is doing. For example, using reference class semantics, $\mathrm{prob}(\psi(\mathbf{e})) > 0$ implies $[\psi(\boldsymbol{v})]_{\boldsymbol{v}} > 0$, whereas with random event semantics we can only derive $\exists \boldsymbol{v}\psi(\boldsymbol{v})$. Only in default models of $\mathrm{prob}(\psi(\mathbf{e})) > 0$ will $[\psi(\boldsymbol{v})]_{\boldsymbol{v}} > 0$ also have to be true (because cross-entropy minimizing measures, in particular, have to be absolutely continuous with respect to the statistical measure).

Another noteworthy phenomenon is that reference class semantics require a sufficiently large domain if subjective probabilities for $\mathbf{e}$ are to differ from the general statistics. For an illustration, consider the following example which is an alternative representation of the die-toss example (cf. example 5.4.13).

$$\forall v(v = \mathsf{r}_1 \dot{\vee} \ldots \dot{\vee} \mathsf{r}_6) \tag{6.13}$$

$$\textstyle\bigwedge_{i=1}^{6} [v = \mathsf{r}_i]_v = \frac{1}{6} \tag{6.14}$$

$$\mathrm{prob}(\mathrm{t} = \mathsf{r}_1 \vee \mathrm{t} = \mathsf{r}_2 \vee \mathrm{t} = \mathsf{r}_3) = 0.3. \tag{6.15}$$

In this encoding the domain of discourse is just the six element set $\{\mathsf{r}_1, \ldots, \mathsf{r}_6\}$ of possible outcomes of a toss of a die (whereas in example 5.4.13 the domain has been thought of as consisting of all throws of arbitrary (fair) dice). The conjunction of (6.13)-(6.15) is satisfiable in random event semantics, but not in reference class semantics because in the set $\{\mathsf{r}_1, \ldots, \mathsf{r}_6\}$, equipped with the uniform probability distribution, there does not exist any subset in which $\{\mathsf{r}_1, \mathsf{r}_2, \mathsf{r}_3\}$ has a conditional probability of 0.3. The encoding used in example 5.4.13, in contrast, is satisfiable in reference class semantics.

On the other hand, every L^{β}-formula ϕ that is satisfiable with respect to reference class semantics also is satisfiable in $\mathscr{L}^{\beta}$: the conditional distribution on S_{e} in a reference class model of ϕ can be used as the belief measure in a feasible model of ϕ.

Reference class semantics, like the random worlds method, do not support default reasoning about probabilities in a way that is independent from making default assumptions about the statistical distribution.

Since this is not the place to conduct an in-depth investigation of a complete first-order version of Paris and Vencovská' s reference class semantics and the logical properties of maximum entropy inference when used in conjunction with this semantics, we here have to leave with a preliminary resumé: in reference class semantics admissible degrees of belief, more than in random event semantics, depend on the structure of the domain and the given statistical information. This makes fewer L^β-formulas satisfiable in reference class semantics than in random event semantics, the latter being in turn more restrictive than possible worlds semantics.

As a basis for performing probabilistic default inferences, the reference class semantics combined with the maximum entropy inference process have more in common with the random worlds method than with $\mathscr{L}^\beta_{\text{def}}$, because in both these methods the derivation of degrees of belief is compounded with making default assumptions about the statistical distribution, which $\mathscr{L}^\beta_{\text{def}}$ has been designed to avoid.

6.3 Conclusion

In this study of default reasoning about probabilities we have been following two main objectives: first, it has been our aim to clarify the epistemic foundations of this form of commonsense reasoning, and to define a general analytical rule by which it should be performed. Secondly, we wanted to devise a logical system in which this reasoning method is formalized.

The first of these two goals has been achieved by analyzing the process of default reasoning about probabilities in terms of thought experiments. For a large class of inference problems concerning the combination of statistical and subjective probabilities – those that are describable in terms of an uncertain event belonging to the domain of possible events described by the statistical information – we have found a suitable epistemic model from which it has been possible to derive the minimum cross-entropy principle as the analytic rule for default reasoning about probabilities.

It then has been shown that the minimum cross-entropy principle can be implemented in the semantics of an extension of full first-order logic for representing statistical and subjective probabilities. Furthermore, we were able to show that both syntax and semantics of the resulting logic $\mathscr{L}^\beta_{\text{def}}$ can be reduced to standard first-order logic.

Thus, we have described a formalism for representing, reasoning, and default reasoning about probabilities that is completely contained within first-order predicate logic. Particularly, the given formalism inherits the completeness of first-order logic. The price we have to pay for this completeness is that probabilities must be allowed to take values in arbitrary real closed fields, not just the real numbers.

The full logic $\mathscr{L}^\beta_{\text{def}}$ is a very rich reference system in which a wide range of default reasoning about probabilities can be performed and analyzed – arguably covering all those cases where default reasoning about probabilities can be understood as a principled application of inference rules with a sound epistemic basis. Inferences in that logic being just as complex as in full first-order logic, one will probably not attempt to implement the unrestricted logic in an automated inference system. Rather, one will look for suitable fragments of $\mathscr{L}^\beta_{\text{def}}$ in which reasoning becomes somewhat more tractable. First important applications of this kind, which have been investigated in [Jaeger, 1994], are probabilistic extensions of terminological logics. These are particularly well-behaved fragments of $\mathscr{L}^\beta_{\text{def}}$ in which we can obtain decidability results even with respect to real-valued probabilities.

List of Symbols

Bibliography

[Abadi and J.Y.Halpern, 1989] M. Abadi and J.Y.Halpern. Decidability and expressiveness for first-order logics of probability. In *Proceedings of the Annual Symposium on Foundations of Computer Science FOCS 30*, 1989.

[Aleliunas, 1990] R. Aleliunas. A new normative theory of probabilistic logic. In H.E. Kyburg et al., editor, *Knowledge Representation and Defeasible Reasoning*. Kluwer Academic Publishers, 1990.

[Bacchus *et al.*, 1992] F. Bacchus, A. Grove, J.Y. Halpern, and D. Koller. From statistics to beliefs. In *Proc. of National Conference on Artificial Intelligence (AAAI-92)*, 1992.

[Bacchus *et al.*, 1993] F. Bacchus, A. Grove, J.Y. Halpern, and D. Koller. Statistical foundations for default reasoning. In *Proc. of International Joint Conference on Artificial Intelligence (IJCAI-93)*, 1993.

[Bacchus *et al.*, 1994a] F. Bacchus, A. Grove, J.Y. Halpern, and D. Koller. Generating new beliefs from old. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Francisco, CA, 1994.

[Bacchus *et al.*, 1994b] F. Bacchus, A. J. Grove, J. Y. Halpern, and D. Koller. From statistical knowledge bases to degrees of belief. Research report, IBM Research Division, 1994.

[Bacchus, 1990a] F. Bacchus. Lp, a logic for representing and reasoning with statistical knowledge. *Computational Intelligence*, 6:209–231, 1990.

[Bacchus, 1990b] F. Bacchus. *Representing and Reasoning With Probabilistic Knowledge*. MIT Press, 1990.

[Bahadur, 1971] R.R. Bahadur. *Some limit theorems in statistics*. CBMS-NSF Regional Conference Series in Applied Mathematics; 4. SIAM, Philadelphia PA, 1971.

[Carnap, 1950] R. Carnap. *Logical Foundations of Probability*. The University of Chicago Press, 1950.

[Cohn, 1993] D. Cohn. *Measure Theory*. Birkhäuser, 1993.

[Csiszár, 1975] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3:146–158, 1975.

[Darwiche and Ginsberg, 1992] Adnan Y. Darwiche and Matthew L. Ginsberg. A symbolic generalization of probability theory. In *Proceedings of National Conference on Artificial Intelligence (AAAI'92)*, 1992.

[de Finetti, 1937] Bruno de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7, 1937. English Translation in [Kyburg and Smokler, 1964].

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

[Diaconis and Zabell, 1982] P. Diaconis and S.L. Zabell. Updating subjective probability. *Journal of the American Statistical Association*, 77(380):822–830, 1982.

[Ebbinghaus *et al.*, 1984] H. D. Ebbinghaus, J. Flum, and W. Thomas. *Mathematical Logic*. Springer-Verlag, 1984.

[Fagin and Halpern, 1991] Ronald Fagin and Joseph Y. Halpern. A new approach to updating beliefs. In Piero P. Bonissone, Max Henrion, Laveen N. Kanal, and John F. Lemmer, editors, *Uncertainty in Artificial Intelligence*. North-Holland, 1991.

[Groeneboom *et al.*, 1979] P. Groeneboom, J. Oosterhoff, and F.H. Ruymgaart. Large deviation theorems for empirical probability measures. *Annals of Probability*, 7(4):553–586, 1979.

[Grove *et al.*, 1992a] A.J. Grove, J.Y. Halpern, and D. Koller. Asymptotic conditional probabilities for first-order logic. In *Proc. 24th ACM Symp. on Theory of Computing*, 1992.

[Grove *et al.*, 1992b] A.J. Grove, J.Y. Halpern, and D. Koller. Random worlds and maximum entropy. In *Proc. 7th IEEE Symp. on Logic in Computer Science*, 1992.

[Halmos, 1950] Paul R. Halmos. *Measure Theory*. Van Nostrad Reinhold Company, 1950.

[Halpern, 1990] J.Y. Halpern. An analysis of first-order logics of probability. *Artificial Intelligence*, 46:311–350, 1990.

[Jaeger, 1994] M. Jaeger. Probabilistic reasoning in terminological logics. In J. Doyle, E. Sandewall, and P. Torasso, editors, *Principles of Knowledge Representation an Reasoning: Proceedings of the Fourth International Conference (KR94)*. Morgan Kaufmann, San Francisco, CA, 1994.

[Jaynes, 1978] E.T. Jaynes. Where do we stand on maximum entropy? In R.D. Levine and M. Tribus, editors, *The Maximum Entropy Formalism*, pages 15–118. MIT Press, 1978.

[Jaynes, 1983] E.T. Jaynes. Concentration of distributions at entropy maxima. In R. D. Rosenkrantz, editor, *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*. D. Reidel Publishing Company, Boston MA, 1983.

[Jeffrey, 1965] R.C. Jeffrey. *The Logic of Decision*. McGraw-Hill, 1965.

[Keisler, 1985] H.J. Keisler. Probability quantifiers. In J. Barwise and S. Feferman, editors, *Model-Theoretic Logics*, pages 509–556. Springer-Verlag, 1985.

[Keynes, 1921] J. M. Keynes. *A Treatise on Probability.* Macmillan, London, 1921.

[Kolmogorov, 1950] A. N. Kolmogorov. *Foundations of the Theory of Probability.* Chelsea Publishing Company, New York, 1950.

[Kraus *et al.*, 1990] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.

[Kullback and Leibler, 1951] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of mathematical statistics*, 22:79–86, 1951.

[Kyburg and Smokler, 1964] H. E. Kyburg and H. E. Smokler, editors. *Studies in Subjective Probability.* John Wiley, 1964.

[Lemmer and Barth, 1982] J. F. Lemmer and S. W. Barth. Efficient minimum information updating for bayesian inferencing in expert systems. In *Proceedings of AAAI 82*, 1982.

[Lifschitz, 1986] V. Lifschitz. On the satisfiability of circumscription. *Artificial Intelligence*, 28:17–27, 1986.

[Martin, 1977] Donald A. Martin. Descriptive set theory: Projective sets. In Jon Barwise, editor, *Handbook of mathematical logic.* Elsevier Science Publishers, 1977.

[McCarthy, 1980] J. McCarthy. Circumscription - a form of non-monotonic reasoning. *Artificial Intelligence*, 13:27–39, 1980.

[Paris and Vencovská, 1989] J.B. Paris and A. Vencovská. On the applicability of maximum entropy to inexact reasoning. *International Journal of Approximate Reasoning*, 3:1–34, 1989.

[Paris and Vencovská, 1990] J.B. Paris and A. Vencovská. A note on the inevitability of maximum entropy. *International Journal of Approximate Reasoning*, 4:183–223, 1990.

[Paris and Vencovská, 1992] J.B. Paris and A. Vencovská. A method for updating that justifies minimum cross entropy. *International Journal of Approximate Reasoning*, 7:1–18, 1992.

[Pearl, 1989] J. Pearl. Probabilistic semantics for nonmonotonic reasoning: A survey. In *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, pages 505–516, 1989.

[Rabin, 1977] Michael O. Rabin. Decidable theories. In Jon Barwise, editor, *Handbook of mathematical logic.* Elsevier Science Publishers, 1977.

[Ramsey, 1931] Frank P. Ramsey. Truth and probability. In *The Foundations of Mathematics and Other Logical Essays.* Kegan Paul, London and Harcourt, Brace and Co., New York, 1931. Reprinted in [Kyburg and Smokler, 1964].

[Reichenbach, 1949] H. Reichenbach. *The Theory of Probability.* University of California Press, 1949.

[Reiter, 1980] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[Ressayre, 1993] J.-P. Ressayre. Integer parts of real closed exponential fields. In P. Clote and J. Krajicek, editors, *Arithmetic, Proof Theory and Computational Complexity*, pages 278–288. Oxford University Press, 1993.

[Sanov, 1957] I. N. Sanov. On the probability of large deviations of random variables (in russian). *Mat. Sbornik N. S.*, 42 (84), 1957. English Translation in: Selected Transl. Math. Statist. Prob 1 (1961).

[Savage, 1954] L. J. Savage. *The Foundations of Statistics.* Wiley, New York, 1954.

[Shafer and Tversky, 1985] G. Shafer and A. Tversky. Languages and designs for probability judgment. *Cognitive Science*, 9:309–339, 1985.

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

[Shore and Johnson, 1980] J.E. Shore and R.W. Johnson. 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, 1980.

[Shore and Johnson, 1981] J.E. Shore and R.W. Johnson. Properties of cross-entropy minimization. *IEEE Transactions on Information Theory*, IT-27(4):472–482, 1981.

[Shore and Johnson, 1983] J.E. Shore and R.W. Johnson. 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, 1983.

[Shore, 1986] J.E. Shore. Relative entropy, probabilistic inference, and ai. In L.N. Kanal and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence.* Elsevier, 1986.

[van Campenhout and Cover, 1981] J. M. van Campenhout and T. M. Cover. Maximum entropy and conditional probability. *IEEE Transactions on Information Theory*, IT-27(4):483–489, 1981.

[von Mises, 1951] R. von Mises. *Wahrscheinlichkeit Statistik und Wahrheit.* Springer, 1951.

[Wen, 1988] W.X. Wen. Analytical and numerical methods for minimum cross entropy problems. Technical Report 88/26, Computer Science, University of Melbourne, 1988.

[Weydert, 1995] E. Weydert. Numeric defaults. In *Proceedings of ECSQARU 95*, 1995.

[Wolter, 1986] H. Wolter. Some remarks on exponential functions in ordered fields. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 32:229–236, 1986.