Self-Referential Probability

Catrin Campbell-Moore

Self-Referential Probability

Catrin Campbell-Moore

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Philosophie
an der Ludwig–Maximilians–Universität
München

vorgelegt von
Catrin Campbell-Moore
aus Cardiff

July 6, 2016

Erstgutachter: Prof. DDr. Hannes Leitgeb
Zweitgutachter: Prof. Dr. Stephan Hartmann

Tag der mündlichen Prüfung: 28. Januar, 2016

Abstract

This thesis focuses on expressively rich languages that can formalise talk about probability. These languages have sentences that say something about probabilities of probabilities, but also sentences that say something about the probability of themselves. For example:

(π)   The probability of the sentence labelled π is not greater than $1/2$.

 Such sentences lead to philosophical and technical challenges. For example seemingly harmless principles, such as an introspection principle:

$$\text{If } \mathsf{P}^\ulcorner \varphi \urcorner = x, \text{ then } \mathsf{P}^\ulcorner \mathsf{P}^\ulcorner \varphi \urcorner = x \urcorner = 1$$

lead to inconsistencies with the axioms of probability in this framework.

This thesis aims to answer two questions relevant to such frameworks, which correspond to the two parts of the thesis: "How can one develop a formal semantics for this framework?" and "What rational constraints are there on an agent once such expressive frameworks are considered?". In this second part we are considering probability as measuring an agent's degrees of belief. In fact that concept of probability will be the motivating one throughout the thesis.

The first chapter of the thesis provides an introduction to the framework. The following four chapters, which make up Part I, focus on the question of how to provide a semantics for this expressively rich framework. In Chapter 2, we discuss some preliminaries and why developing semantics for such a framework is challenging. We will generally base our semantics on certain possible world structures that we call probabilistic modal structures. These immediately allow for a definition of a natural semantics in restrictive languages but not in the expressively rich languages that this thesis focuses on. The chapter also presents an overview of the strategy that will be used throughout this part of the thesis: we will generalise theories and semantics developed for the liar paradox, which is the sentence:

(λ)   The sentence labelled λ is not true.

In Chapter 3, we will present a semantics that generalises a very influential theory of truth: a Kripke-style theory (Kripke, 1975) using a strong Kleene evaluation scheme. A feature of this semantics is that we can understand it as assigning sentences intervals as probability values instead of single numbers. Certain axioms of probability have to be dropped, for example $\mathsf{P}_{=1}^\ulcorner \lambda \vee \neg \lambda \urcorner$ is not satisfied in the construction, but the semantics can be seen as assigning

non-classical probabilities. This semantics allows one to further understand the languages, for example the conflict with introspection, where one can see that the appropriate way to express the principle of introspection in this case is in fact to use a truth predicate in its formulation. This follows a strategy from Stern (2014a,b). We also develop an axiomatic system and show that it is complete in the presence of the ω-rule which allows one to fix the standard model of arithmetic.

In Chapter 4, we will consider another Kripke-style semantics but now based on a supervaluational evaluation scheme. This variation is particularly interesting because it bears a close relationship to imprecise probabilities where agents' credal states are taken to be sets of probability functions. In this chapter, we will also consider how to use this language to describe imprecise agents reasoning about one another. These considerations provide us with an argument for using imprecise probabilities that is very different from traditional justifications: by allowing agents to have imprecise probabilities one can easily extend a semantics to languages with sentences that talk about their own probability, whereas the traditional precise probabilist cannot directly apply his semantics to such languages.

In Chapter 5, a revision theory of probability will be developed. In this one retains classical logic and traditional probability theory but the price to pay is that one obtains a transfinite sequence of interpretations of the language and identifying any particular interpretation as "correct" is problematic. In developing this we are particularly interested in finding limit stage interpretations that can themselves be used as good models for probability and truth. We will require that the limit stages "sum up" the previous stages, understood in a strong way. In this chapter two strategies for defining the successor stages are discussed. We first discuss defining (successor) probabilities by considering relative frequencies in the revision sequence up to that stage, extending ideas from Leitgeb (2012). The second strategy is to base the construction on a probabilistic modal structure and use the accessibility measure from that to determine the interpretation of probability. That concludes Part I and the development of semantics.

In Part II, we consider rationality requirements on agents who have beliefs about self-referential probability sentences like π. For such sentences, a choice of the agent's credences will affect which worlds are possible. Caie (2013) has argued that the accuracy and Dutch book arguments should be modified because the agent should only care about her inaccuracy or payoffs in the world(s) that could be actual if she adopted the considered credences. We consider this suggestion for the accuracy argument in Chapter 7 and the Dutch book argument in Chapter 8. Chapter 6 acts as an introduction to these considerations. We will show that these modified accuracy and Dutch book criteria lead to an agent being rationally required to be probabilistically incoherent, have negative credences, fail to be introspective and fail to assign the same credence to logically equivalent sentences. We will also show that this accuracy criterion depends on how inaccuracy is measured and that the accuracy criterion differs from the Dutch book criterion. We will in fact suggest rejecting Caie's suggested modifications. For the accuracy argument, we suggest in Section 7.3 that the agent should consider how accurate the considered credences are from the perspective of her current credences. We will also consider how to generalise this version of the accuracy criterion and present ideas suggesting that it connects to the

semantics developed in Part I. For the Dutch book argument, in Section 8.6 we suggest that this is a case where an agent should not bet with his credences.

We finish the thesis with a conclusion chapter in Chapter 9.

Preface

I am very grateful to my supervisor, Hannes Leitgeb, who gave me the original idea and motivation to start working on this project as well as a lot of help throughout my time as a doctoral student in Munich. A huge amount of thanks goes to Johannes Stern for his never-ending support and for many discussions about this work. The Munich Center for Mathematical Philosophy provided a very stimulating environment for me and I am grateful to everyone who was a part of that. This includes Martin Fischer, Thomas Schindler, Lavinia Picollo, Rossella Marrano, Marta Sznajder, Gil Sagi and Branden Fitelson. Particular thanks goes to Seamus Bradley and the imprecise probabilities/formal epistemology reading group including Conor Mayo-Wilson, Aidan Lyon and Greg Wheeler. During my PhD I spent two months in Buenos Aires and I am very grateful to Eduardo Barrio and the Buenos Aires logic group for making that a very helpful period. I would also like to thank a number of other people who have provided me (often very detailed) comments on my papers, including Michael Caie, Anupam Das, Daniel Hoek, Leon Horsten, Karl-Georg Niebergall, Ignacio Ojea, Richard Pettigrew, Stanislav Speranski and Sean Walsh, as well as some anonymous referees. My work has been greatly helped by insightful questions and comments at talks I have given at Munich, Groningen, Amsterdam, Bristol, Buenos Aires, Oxford, London, Pasadena, Venice, Los Angeles, Manchester, Cambridge and New Brunswick. I am very grateful to the attendees and organisers of these conferences. I would also like to thank Volker Halbach for supervising me before I started my PhD and teaching me much of what I know about truth and everyone who was part of the Philosophy of Maths seminar at Oxford. Thanks also to Robin Knight for teaching me how to do research, and to Brian King and Stephan Williams for introducing me to philosophy and teaching me how to love wisdom. Last, but certainly not least, thank you very much to my family, Mum, Dad, Grandma and Owen, and all my friends!

Chapter 3 is a expansion of the paper Campbell-Moore (2015a), with additional results in Sections 3.3 and 3.4 and a much expanded proof of the completeness theorem in Section 3.5.2. Also parts of Chapter 1, particularly Section 1.1, is a development of the introduction to that paper.

Part II has developed from Campbell-Moore (2015b).

Contents

II Rationality Requirements 157

Nomenclature

$\mathfrak{M}$ A probabilistic modal structure.

M   A model of the base language $\mathcal{L}$ (usually this does not contain P or T).

M   A collection of models of the base language $\mathcal{L}$, so for each world in the probabilistic modal structure, w, $\mathbf{M}(w)$ is itself a model of the base language $\mathcal{L}$.

$\mathcal{M}$ A model of the language including probability (and perhaps truth). Takes the form (M, p).

$\boldsymbol{\mathcal{M}}$ A collection of models for the language including probability (and perhaps truth). One for each world, w, in the probabilistic modal structure. So $\boldsymbol{\mathcal{M}}(w)$ is a model of the expanded language.

P The symbol in the formal language which represents probability. Often a function symbol, sometimes given by predicates $\mathsf{P}_{\geqslant}$ or $\mathsf{P}_{\geqslant r}$.

$\mathbb{P}$ An operator in a formal language representing probability. This modifies a sentence to produce a new sentence which says something about the probability of the original sentence.

p   A function from Sent to $\mathbb{R}$, often assumed to be probabilistic. Used as the interpretation of the object-level P.

p   A collection of functions from Sent to $\mathbb{R}$, one for each world, w, of the probabilistic modal structure. Called a 'prob-eval function'.

c, b An agent's credences in a fixed agenda, $\mathcal{A}$. Formally: a function from $\mathcal{A}$ to $[0, 1]$. Similar to p. We will also use $\mathbf{c}, \mathbf{b}$ for the versions of $\mathbf{p}$ restricted to $\mathcal{A}$.

T The symbol in the formal language representing truth.

$\mathbb{T}$ An operator in a formal language representing truth. Analogous to $\mathbb{P}$.

T   A set of sentences, often maximally consistent; the interpretation of T. Analogous to p.

T   A collection of Ts, one for each world in a probabilistic modal structure. So $\mathbf{T}(w)$ is a (usually maximally consistent) set of sentences. Analogous to $\mathbf{p}$.

f Essentially works like **T**, except where $f(w)$ is a set of *codes of* sentences. Called an 'evaluation function'.

w A "world" in the probabilistic modal structure.

w   A model of the language including probability, restricted to a specific set of sentences $\mathcal{A}$. For use in arguments for rationality requirements.

Chapter 1

Introduction

1.1 What are we interested in and why?

This thesis will study frameworks where there are sentences that can talk about their own probabilities. For example they can express the sentence π:

(π)   The probability of π is not greater than or equal to $1/2$.

Consider the following empirical situation that displays features similar to π,[1] which is a modification of an example by Greaves (2013):

> Alice is up for promotion. Her boss, however, is a deeply insecure type: he will only promote Alice if she comes across as lacking in confidence. Furthermore, Alice is useless at play-acting, so she will come across that way iff she really does have a low degree of confidence that she'll get the promotion. Specifically, she will get the promotion exactly if she does not have a degree of belief greater than or equal to $1/2$ that she will get the promotion.

This is a description of a situation where a sentence, *Promotion*, is true just if her degree of belief in that very sentence satisfies some property. This is the same as for π.

Such languages can be problematic. For example, contradictions can arise between seemingly harmless principles such as probabilism and introspection.

A possible response to this is to circumvent such worries by preventing such sentences from appearing in the language. However, we shall argue that the result of doing that is that one cannot properly represent quantification or formalise many natural language assertions or interesting situations, so we think that is the wrong path to take. Instead we will suggest that such self-referential probability assertions should be expressible, but one should work out how to deal with this language and how to circumvent such contradictions. In this thesis we will do just that.

One important aspect of that is to develop *semantics* which tell us when such self-referential sentences are true or not. This is what we will do in Part I. The sentence π bears a close relationship to the liar paradox, which is a sentence that says it is not true, and many of our considerations will bear close relationships to considerations from the liar paradox.

---

[1]A discussion of how *Promotion* and π connect can be found on Page 6.

1.1.1 Why should self-referential probability sentences be expressible?

Probability and probabilistic methods are heavily used in many disciplines, including, increasingly, philosophy. We will consider formal languages that can talk about probability, so we will assume that they can formalise at least simple expressions about probability such as:

The probability of the coin landing heads is $1/2$.

We will work with a framework where probabilities are assigned to *sentences* instead of to events, which are subsets of a sample space. Although this is uncommon in mathematical study of probability, it is not uncommon in philosophical work and will allow us to develop logics for probability. This is also the approach that is often taken in computer science. We would then state the axioms of probability sententially, so for example have the axiom:

If φ is a logical tautology, then $\mathrm{p}(\varphi) = 1$.

There is typically a correspondence between probabilities assigned to sentences and those assigned to events that are subsets of a space of all possible models,[2] with the correspondence

$$\mathrm{p}(\varphi) = m\{\mathcal{M} \in \mathsf{Mod} \mid \mathcal{M} \models \varphi\}.$$

In fact, for the work here it is not important that we assign probabilities to *sentences*, instead it is important that the objects to which the probabilities are assigned have sufficient syntactic-style structure. The sufficient structure that we require is that operations analogous to syntactic operations can be defined on the objects.[3] Events, or subsets of the sample space, do not have this sufficient structure. For simplicity, in this thesis we will assume that probabilities are assigned to sentences. In the formal language we will in fact have that probabilities are attached to natural numbers that act as codes of sentences.

Express higher order probabilities and quantification

We want to be able to express embeddings of probabilities, as this is useful to express relationships between different notions of probability. Consider the example from Gaifman (1988) who takes an example from the New Yorker of a forecaster making the following announcement:

> There is now 60% chance of rain tomorrow, but, there is 70% chance that later this evening the chance of rain tomorrow will be 80%.

This expresses a fact about the current chances of rain, $\mathsf{ch}_{t_0}(\ulcorner Rain_{t_2} \urcorner) = 0.6$, as well as a fact about the current chances of some later chances,

$$\mathsf{ch}_{t_0}(\ulcorner \mathsf{ch}_{t_1}(\ulcorner Rain_{t_2} \urcorner) = 0.8 \urcorner) = 0.7.$$

Other cases where one kind of probability is assigned to another can be found in Lewis's Principal Principle, (Lewis, 1980), which says that an agent should

---

[2] See Section 1.2.1.

[3] See Halbach (2014, ch. 2) for more information about this.

defer to the objective chances. A bit more carefully, it says: conditional on that the objective chance of A is r, one should set ones subjective credence in A to also be r. We can then formulate this principle, for an agent, Tom, as:

$$\mathsf{P}^{\mathrm{Tom}}(A \mid \ulcorner \mathsf{ch}(A) = r \urcorner) = r$$

Such embedded probabilities are also required to express agents' beliefs about other agents' beliefs. For example:

Georgie is (probabilistically) certain that Dan believes to degree $1/2$ that the coin will land heads.

which can be formulated as:

$$\mathsf{P}^{\mathrm{Georgie}}(\ulcorner \mathsf{P}^{\mathrm{Dan}}(\ulcorner Heads \urcorner) = 1/2 \urcorner) = 1.$$

Since we are working in a sentential framework, we can express this sentence and talk about it without first being aware of exactly when $\mathsf{P}^{\mathrm{Dan}}(\ulcorner Heads \urcorner) = 1/2$ is true, i.e. which set of worlds, or event, it corresponds to. This is particularly important because determining which set of worlds a sentence corresponds to is essentially to give a semantics, but we will see that developing semantics for self-referential probabilities is not simple, so we shouldn't build into the framework that we already know which set of worlds it corresponds to.

We will therefore consider languages which can express such embedded probabilities,[4] so will allow constructions of the form:

$$\mathsf{P}^A(\ldots \mathsf{P}^B(\ldots \mathsf{P}^A \ldots) \ldots) \ldots$$

We will furthermore allow for self-applied probability notions, or higher order probabilities, namely constructions such as

$$\mathsf{P}^A(\ldots \mathsf{P}^A \ldots) \ldots$$

These offer us two advantages. Firstly, they allow for a systematic syntax once one wishes to allow for embedded probabilities, as one then does not need to impose syntactic restrictions on the formulation of the language. Secondly, their inclusion may be fruitful, as was argued for in Skyrms (1980). For example, we can then formalise facts about the introspective abilities of an agent, or the uncertainty or vagueness about the first order probabilities. One might disagree and argue that they are trivial and collapse to the first level, however even then one should still allow such sentences to be expressed in the language and instead include an extra principle to state this triviality of the higher levels. Such a principle would be an introspection principle, which is a formalisation of:

---

[4]Gaifman (1988) assigns probabilities to members of an event space $\mathcal{F} \subseteq \wp(\Omega)$ and he makes the assumption on this that there is an operation PR with

$$\mathsf{PR} : \mathcal{F} \times \text{ set of closed intervals } \to \mathcal{F}$$

satisfying certain properties. So for each event $A \in \mathcal{F}$ and interval Δ there is an event that, informally, says that the true probability of A lies in the interval Δ. (We will relax this interpretation and be interested in general relationships between probabilities, for example in one agent's attitudes towards another agent's attitudes.) It is much easier to make the assumption that we have such embedded probabilities in the sentential framework than in the events framework because because in the events framework it is required that $\mathsf{PR}(A, \Delta) \subseteq \Omega$ but there is no analogous assumption for the sentential variety. We can define the language first and then determine a semantics afterwards.

> If the probability of φ is $\geqslant r$,
> then the probability of "The probability of φ is $\geqslant r$" is 1.

In fact we will see that this principle will lead to challenges once self-referential sentences like π are allowed for.

There are two main ways of giving languages that can express higher order probabilities but do not allow for self-referential probabilities. The first is to consider a hierarchy of languages or a typed probability notion. This is given by a language $\mathcal{L}_0$ that cannot talk about probabilities at all, together with a metalanguage $\mathcal{L}_1$ that can talk about probabilities of the sentences of $\mathcal{L}_0$, together with another metalanguage $\mathcal{L}_2$ that can talk about the probabilities of sentences of $\mathcal{L}_1$ and $\mathcal{L}_0$, etc. This leads to a sequence of languages $\mathcal{L}_0, \mathcal{L}_1, \mathcal{L}_2, \ldots$ each talking about probabilities of the previous languages. In ordinary language we can talk about multiple probability notions, such as objective chance and the degrees of beliefs of different agents, but the different notions should be able to apply to all the sentences of our language and there should not be a hierarchy of objective chance notions $\mathsf{ch}_0, \mathsf{ch}_1, \ldots$ applying to the different levels of language. However that is what one would obtain by having the idea of a hierarchy of languages, each containing their own probability notions that apply to the previous languages. This would be the approach that corresponds to Tarski's proposal in response to the liar paradox (Tarski, 1956). Tarski's solution has been rejected for various reasons including the one analogous to that just suggested: in ordinary language we don't have a hierarchy of truth predicates, but instead we have one truth predicate that can apply to all sentences of the language (Kripke, 1975).

The second approach is to instead consider one language where the probability notion is formalised by an operator. This is the approach taken in Aumann (1999); Fagin et al. (1990); Ognjanović and Rašković (1996); Bacchus (1990), amongst others. Each of these differ in their exact set-up but the idea is that one adds a recursive clause saying: if φ is a sentence of the language then we can form another sentence of the language that talks about the probability of φ. For example in Aumann (1999) and Ognjanović and Rašković (1996), one adds the clause:

$$\text{If } \varphi \in \mathcal{L} \text{ then } \mathbb{P}_{\geqslant r}\varphi \in \mathcal{L}$$

to the construction rules of the language $\mathcal{L}$.[5] In this language $\mathbb{P}_{\geqslant r}$ acts syntactically as an operator like $\neg$ instead of like a predicate so this is not a language of first order logic but is instead an operator logic.

Both the typed and operator languages avoid self-referential probabilities, but they cannot easily account for quantification over all of the sentences of the language. So for example they cannot express:

All tautologies have probability 1. (1.1)

Artemisa is certain that Chris has some non-extremal degrees of belief. (1.2)

The former is an axiom of probability theory, so something we would like to be able to write as an axiom in a formal language. Although this could be expressed by a schema, i.e. a collection of sentences of the language, existential

---

[5] For some choice of values of r, for example the rational numbers.

4

sentences like, "Chris has some non-extremal degrees of belief", could not. Sentence (1.2) is a statement of the interaction between two agents that we may want to be able to express and discuss.

There is a language for reasoning about probabilities that *can* express this quantification: one can formalise probability within standard first order logic by either adding predicate symbols, $\mathsf{P}_{\geqslant r}$, or a function symbol, P. To have this appropriately formalise probability, one needs to have the objects to which one assigns probabilities, or representations thereof, in the domain of the first order model. Typically this is done by assuming that we have available a background of arithmetic and a way of coding the sentences of the language into numbers, usually via a so-called Gödel coding, so for each sentence φ there will be some natural number $\#\varphi$ which represents it, and the formal language will refer to this by $\ulcorner \varphi \urcorner$. For a formal introduction to this see Section 1.6. These are the kinds of languages that we study in this thesis.

In these languages we can now easily formulate quantified claims like (1.1) and (1.2) just by using usual first order quantification, e.g. by:

$$\forall x(\mathsf{Prov}(x) \to \mathsf{P}_{=1}(x)) \tag{1.3}$$

$$\mathsf{P}^{\mathrm{Artemisa}}_{=1}\ulcorner \exists x(\mathsf{P}^{\mathrm{Chris}}_{>0}(x) \wedge \mathsf{P}^{\mathrm{Chris}}_{<1}(x))\urcorner \tag{1.4}$$

If one takes enough arithmetic as a background theory[6] then one can derive the diagonal lemma for this language and therefore result in admitting sentences that talk about their own probabilities. So by formalising probability as a (type-free) predicate, we can prove that sentences like π must exist. More carefully, we can then see that there is a sentence, called π, where

$$\pi \leftrightarrow \neg \mathsf{P}_{\geqslant 1/2}\ulcorner \pi \urcorner$$

is arithmetically derivable. This is analogous to the situation in Gödel's incompleteness theorem where one shows that there must be a Gödel sentence G such that

$$G \leftrightarrow \neg \mathsf{Prov}_{\mathsf{PA}}\ulcorner G \urcorner$$

is arithmetically derivable. Such self-referential probabilities therefore arise when we consider languages that can express such quantification. This ability to express quantification is a very persuasive argument in favour of the predicate approach to probabilities.

One might try to instead account for this quantification by working with operators and *propositional quantification*. However, if we give propositional quantification a substitutional understanding it quickly becomes equivalent to a truth predicate and self-reference becomes expressible (this argument is presented in Halbach et al., 2003; Stern, 2015b).

There is a further expressive resource which we may wish to obtain in our formal language and which will result in self-referential probabilities: the ability to refer back to expressions and talk about substitutions of these expressions This is what happens when we say:

(π)   The probability of the sentence labelled π is not greater than $1/2$.

So having this ability will result in obtaining self-referential probability sentences. Further, this is an expressive resource that is fundamental to natural

---

[6]Peano arithmetic will suffice. See Section 1.6.

language so should also be available in a formal language. This line of argument is taken from Stern (2015b, sec. 2.3.3).

So to sum up this discussion here: the idea is that if we want to be able to have a formal language that can express higher order probabilities and have the ability to express quantification into the probability notion, then such self-referential probabilities end up being expressible.

What we can now express

Self-referential probabilities might themselves be useful to express certain situations.

The example of Alice and the promotion from Page 1 was a situation which expresses self-reference. The sentence "Alice will get the promotion" is true if and only if she has a low degree of belief in that very sentence. In the expressively rich languages there is already a sentence with that feature, π, which can be used to interpret Alice's situation. In the operator language we would add an additional axiom $Promotion \leftrightarrow \neg \mathbb{P}_{\geqslant 1/2} Promotion$. What is ultimately important in our considerations of the language we should use is not whether it is a predicate or operator language but instead whether it has such diagonal sentences around. If one adds by force diagonal sentences to the operator language, for example to represent Alice's situation, then we are in the same ball-park as when we consider predicate languages that already have such expressive power inbuilt (because of the Diagonal Lemma) and the same problems will arise.[7] So if one thinks that situations like Alice's are possible, one will be forced to worry about these issues. By dealing with a framework that can already express all these diagonal sentences (which we do by working with a first order language for formulating probability) we obtain a framework where any (self-referential) set-up can be easily considered. Of course, there is a place for studying restrictive frameworks, but the step to the expressive frameworks is important as they can describe situations that might arise.

One might think that π doesn't appropriately formalise Alice's situation because Alice's situation doesn't really express self reference. Alice might just be unaware of her boss's intentions then there is no problem, she can have some credences in $Promotion$ without knowing that that affects the truth value of $Promotion$. For π, the equivalence between π and $\neg \mathsf{P}_{\geqslant 1/2} \ulcorner \pi \urcorner$ holds whenever arithmetic holds, which we will assume to hold everywhere. For $Promotion$, the equivalence is imposed on the setup. If it is common knowledge[8] that the equivalence holds, i.e. where all agents are certain that all agents are certain . . . that $Promotion \leftrightarrow \neg \mathsf{P}_{\geqslant 1/2}^{\text{Alice}} \ulcorner Promotion \urcorner$, then we can model the situation by only considering worlds where the equivalence holds. If that assumption is made then π can appropriately formalise $Promotion$.

Egan and Elga (2005) argued that in a situation like Alice's, she should not believe the equivalence between $Promotion$ and $\neg \mathsf{P}_{\geqslant 1/2}^{\text{Alice}} \ulcorner Promotion \urcorner$. They therefore say that although it seems like Alice should have learned the equivalence, that is in fact not the case since rational agents should never learn that they are anti-experts, i.e. that their (all-things-considered) attitudes anti-corrolate with the truth. Their argument is based on the inconsistency of believing this equivalence, being introspective and being probabilistically coherent.

---

[7] See Stern (2015b) and references therein for Diagonal Modal Logic.

[8] By which we mean common probabilistic certainty.

Such a response isn't available in the case of π, so we have to account for situations where the agent does believe that she is an anti-expert about π. Once we then have to study that, we can also use these considerations for *Promotion* and allow an agent to learn that she is an anti-expert about *Promotion* too.

Here is another example in a similar spirit, now modified from Carr (ms):

> Suppose your (perfectly reliable) yoga teacher has informed you that the only thing that could inhibit your ability to do a handstand is self-doubt, which can make you unstable or even hamper your ability to kick up into the upside-down position. In fact you will be able to do a handstand just if you believe you will manage to do it to degree greater than (or equal to) a half.

This is a situation where

$$Handstand \leftrightarrow \mathsf{P}_{\geqslant 1/2}\ulcorner Handstand \urcorner$$

is stipulated to be true (and common knowledge). This situation is analogous to the case of Alice and her promotion except it is now not an *undermining* sentence but a *self-supporting* sentence.

There is a further way that one might come across self-referential sentences in natural language or natural situations: In natural language we can assert sentences that are self-referential or not depending on the empirical situation and an appropriate formal language representing natural language should be able to do this too.[9]

Consider the following example.

> Suppose that Smith is a Prime Ministerial candidate and the candidates are campaigning hard today. Smith might say:
>
> > I don't have high credence in anything that the man who will be Prime Minister says today. $\qquad$ (1.5)
>
> Imagine further, that unknown to Smith, he will win the election and will become Prime Minister.

Due to the empirical situation, (1.5) expresses a self-referential probability assertion analogous to π, the self-referential probability example from Page 1. To reject Smith's assertion of (1.5) as formalisable would put serious restrictions on the natural language sentences that are formalisable.

Truth as a predicate

Such discussions are not new. The possibility of self-reference is also at the heart of the liar paradox, namely a sentence that says of itself that it is not true. This can be expressed by:

(λ) $\quad \lambda$ is not true

In Kripke's seminal paper he says:

---

[9]An example of empirical self-reference in the case of truth is Kripke's Nixon example from Kripke (1975, pg. 695).

> Many, probably most, of our ordinary assertions about truth or falsity are liable, if our empirical facts are extremely unfavourable, to exhibit paradoxical features.... it would be fruitless to look for an intrinsic criterion that will enable us to sieve out—as meaningless, or ill-formed—those sentences which lead to paradox. (Kripke, 1975, p. 691–692)

Analogously, if we wish our formal language to represent our ordinary assertions about probability, for example the case of Smith and his distrust in the prime ministerial candidate, we should allow for the possibility of self-referential sentences. We should then provide a clear syntax and semantics that can appropriately deal with these sentences as well as providing an axiomatic theory for reasoning about the language. This is one of the important goals of this thesis, and it is what we will work on in Part I.

The fact that truth is usually understood to be a predicate leads us to a further argument for understanding probability as a predicate: if possible, different notions, like truth and probability, should be formalised in a similar way. This argument has also been used for formalising (all-or-nothing) modalities as predicates.

> It seems arbitrary that some notions are formalised as predicates, while others are conceived as operators only. It also forces one to switch between usual first-order quantification and substitutional quantification without any real need. (Halbach et al., 2003, p. 181)

There has been work on arguing that modalities, such as belief, knowledge or necessity, should be formulated as predicates, and these arguments will often also apply to our setting where we suggest that *probability* should be formalised by first order logic means, either by a predicate or function symbol.

Stress-test

Such self-referential probabilities are also interesting because they can provide a stress test for accounts of probability. If an account or theory of probability also works when self-referential probabilities are admitted into the framework, then this shows that the theory is robust and can stand this stress test.

For example, in Caie (2013) such self-referential probabilities have recently been used to argue that traditional analysis of rationality requirements on agents does not appropriately apply to self-referential probabilities and that in such a setting an agent may be rationally required to be probabilistically incoherent. We will further study Caie's proposal in Part II and reject his modification.

1.1.2   Some previous work on self-referential probabilities

In Leitgeb (2012), Leitgeb develops the beginnings of what might be called a revision semantics for probability, though he only goes to stage ω. He also provides a corresponding axiomatic theory. Our work in Chapter 5 can be seen as an extension of the work in that paper.

In Caie (2013) and Caie (2014), Caie argues that traditional arguments for probabilism, such as the argument from accuracy, the Dutch Book argument and the argument from calibration, all need to be modified in the presence of

self-referential probabilities, and that so modified they do not lead to the rational requirement for beliefs to be probabilistic. In Part II, which is in part a development of Campbell-Moore (2015b), we more carefully consider Caie's suggested modifications of the rational requirements from accuracy and Dutch book considerations. In Caie (2013), Caie also presents a prima facie argument against probabilism by noticing its inconsistency with the formalised principles of introspection when such self-reference is present. We will argue in this thesis that one should reformulate an introspection principle by using a truth predicate. Once one adopts a consistent theory of truth this will avoid inconsistency. This is discussed in sections 3.4.1, 1.4 and 2.4.2. Initial work by Caie on this topic is in his thesis, Caie (2011). There (p. 63) he provides a strong Kleene Kripke style semantics which could be seen as a special case of our semantics developed in Chapter 3,[10] though I was unaware of his construction before developing the version here.

A probabilistic liar is also mentioned in Walsh (2013) who notes it can cause a problem for a probabilistic account of confirmation, though leaves analysis of this problem to future work.

Lastly, the unpublished paper Christiano et al. (ms) also considers the challenge that probabilism is inconsistent with introspection. In their paper Christiano et al. show that probabilism is consistent with an approximate version of introspection where one can only apply introspection to open intervals of values in which the probability lies. Their result is interesting, but in this thesis we will discuss the alternative suggestion just mentioned that introspection should be reformulated. These authors come from a computer science background and believe that these self-referential probabilities might have a role to play in the development of artificial intelligence.

Although there is not yet much work on self-referential probabilities, there is a lot of work in closely related areas. There is a large amount of research into theories of truth, which we will draw from for this thesis. Much of that setup and framework is presented in (Halbach, 2014). There is also existing work on the predicate approach to modality. Papers I have used heavily in the writing of thesis are Halbach et al. (2003); Halbach and Welch (2009); Stern (2015b, 2014a,b, 2015a), where these authors consider frameworks with sentences that can talk about their own necessity and use possible world framework to analyse them. The semantics that we study that are based on probabilistic modal structures can be seen as generalisations of their semantics.

1.2 What is the notion of probability for our purposes?

In this thesis we are trying to capture the notion of probability in an expressively rich language. Before we can get started on that we should carefully introduce a formal notion of probability and then discuss interpretations of it.

---

[10]His semantics works with a single probability measure over W instead of measures for each w, and his definitions don't give the equivalent of the non-consistent evaluation functions.

1.2.1 Probability axioms

We will talk about probability in two different, but related, ways. Firstly the set-up that is common in the mathematical study of probability theory is to consider probabilities attaching to events, or subsets of some sample space.

Definition 1.2.1. A *probability space* is some $\langle \Omega, \mathcal{F}, m \rangle$ such that:

- Ω is a non-empty set, called the *sample space*.

- $\mathcal{F}$ is a *Boolean algebra over* Ω, i.e. $\mathcal{F} \subseteq \wp(\Omega)$ such that:
    - $\varnothing \in \mathcal{F}$, $\Omega \in \mathcal{F}$,
    - If $A \in \mathcal{F}$ then $\Omega \setminus A \in \mathcal{F}$,
    - If $A, C \in \mathcal{F}$, then $A \cup C \in \mathcal{F}$
    - We call $\mathcal{F}$ a σ-*algebra* if it also satisfies: for any countable collection $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathcal{F}$, $\bigcup_{i \in \mathbb{N}} A_i \in \mathcal{F}$.

- $m : \mathcal{F} \to \mathbb{R}$ is a *finitely additive probability measure* (over $\mathcal{F}$), i.e.:
    - For all $A \in \mathcal{F}$, $m(A) \geqslant 0$
    - $m(\Omega) = 1$
    - If $A, C \in \mathcal{F}$ and $A \cap C = \varnothing$, then $m(A \cup C) = m(A) + m(C)$
    - We call m *countably additive* if it also satisfies: for any countable collection $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathcal{F}$ of pairwise disjoint sets (i.e. for $i \neq j$, $A_i \cap A_j = \varnothing$) if $\bigcup_{i \in \mathbb{N}} A_i \in \mathcal{F}$ then $m(\bigcup_{i \in \mathbb{N}} A_i) = \sum_{i \in \mathbb{N}} m(A_i)$.[11]

If m is not countably additive we may call it a *merely finitely additive probability measure*.

In fact we will typically be considering probability spaces where $\mathcal{F} = \wp(\Omega)$ and where m is not assumed to be countably additive. This will generally be possible because of the following well-known theorem:

Proposition 1.2.2. *If* $\langle \Omega, \mathcal{F}, m \rangle$ *is a probability space, then for any Boolean algebra* $\mathcal{F}^* \supseteq \mathcal{F}$*, there is an extension of* m *to* m^* *which is a finitely additive probability measure over* $\mathcal{F}^*$*. In particular, it can always be extended to* $\wp(\Omega)$*.*

We are generally interested in considering a logic describing probability and as discussed in the introduction, for that purpose we will generally consider p as a function assigning real number to *sentences*. We then define what it is for some such p to be probabilistic. In fact this is what is often considered in logic and philosophy.

Definition 1.2.3. Let $\mathsf{Mod}_{\mathcal{L}}$ be the collection of all models of $\mathcal{L}$.

p : $\mathsf{Sent}_{\mathcal{L}} \to \mathbb{R}$ is *probabilistically coherent*, or *probabilistic*, iff for all $\varphi, \psi \in \mathsf{Sent}_{\mathcal{L}}$,

- If $\mathcal{M} \models \varphi$ for all $\mathcal{M} \in \mathsf{Mod}_{\mathcal{L}}$ then $\mathrm{p}(\varphi) = 1$

- $\mathrm{p}(\varphi) \geqslant 0$

---

[11] By the Caratheodory extension theorem, such an m can always be extended to a countably additive measure over the smallest σ-algebra extending $\mathcal{F}$.

- If $\mathcal{M} \models \neg(\varphi \wedge \psi)$ for all $\mathcal{M} \in \mathsf{Mod}_{\mathcal{L}}$, then

$$\mathrm{p}(\varphi \vee \psi) = \mathrm{p}(\varphi) + \mathrm{p}(\psi)$$

Suppose $\mathcal{L}$ is a language extending the language of (Peano-)arithmetic, (L_{PA}, see Definition 1.6.1). p is $\mathbb{N}$-*additive* iff[12]

$$\mathrm{p}(\exists x \varphi(x)) = \lim_{n} \mathrm{p}(\varphi(\overline{0}) \vee \ldots \vee \varphi(\overline{n})).$$

We say p is *probabilistic over a theory* Γ if whenever $\Gamma \vdash \varphi$, $\mathrm{p}(\varphi) = 1$. This is equivalent to replacing $\mathsf{Mod}_{\mathcal{L}}$ in the definition by all models of the language $\mathcal{L}$ satisfying the theory Γ, $\mathsf{Mod}_{\mathcal{L}}^{\Gamma}$.

This definition of p being probabilistic is essentially saying that it is given by a probability measure over the models.

Proposition 1.2.4. $\mathrm{p} : \mathsf{Sent}_{\mathcal{L}} \to \mathbb{R}$ *is probabilistic iff there is some m which is a finitely additive probability measure over* $\langle \mathsf{Mod}_{\mathcal{L}}, \wp(\mathsf{Mod}_{\mathcal{L}}) \rangle$ *such that*

$$\mathrm{p}(\varphi) = m\{\mathcal{M} \in \mathsf{Mod}_{\mathcal{L}} \mid \mathcal{M} \models \varphi\}.$$

$\mathrm{p} : \mathsf{Sent}_{\mathcal{L}} \to \mathbb{R}$ *is probabilistic over* Γ *if there is some m a finitely additive probability measure over* $\langle \mathsf{Mod}_{\mathcal{L}}^{\Gamma}, \wp(\mathsf{Mod}_{\mathcal{L}}^{\Gamma}) \rangle$ *such that*

$$\mathrm{p}(\varphi) = m\{\mathcal{M} \in \mathsf{Mod}_{\mathcal{L}}^{\Gamma} \mid \mathcal{M} \models \varphi\},$$

where $\mathsf{Mod}_{\mathcal{L}}^{\Gamma}$ *denotes the set of all $\mathcal{L}$-models of the theory Γ.*

Being $\mathbb{N}$-additive in part captures the notion that m is countably additive, but it also captures the idea that the domain is exactly $0, 1, 2 \ldots$. It was called σ-additivity in Leitgeb (2008) and Leitgeb (2012). For the case where the language is just L_{PA}, this is the Gaifman condition (see, e.g. Scott and Krauss, 2000). We have a corresponding result for $\mathbb{N}$-additivity, which makes precise our claim that $\mathbb{N}$-additivity states σ-additivity and fixes the domain to be $\mathbb{N}$.

Definition 1.2.5. Suppose $\mathcal{L}$ is a language extending L_{PA}.

We say that $\mathcal{M}$ is an $\mathbb{N}$-*model* if the L_{PA}-reduct of $\mathcal{M}$ is the standard natural number structure, $\mathbb{N}$, i.e. if $\mathcal{M}$ interprets the arithmetic vocabulary as in the standard model of arithmetic and has a domain consisting of the collection of standard natural numbers.

Let $\mathsf{Mod}_{\mathcal{L}}^{\mathbb{N}}$ denote the collection of $\mathbb{N}$-models.

Proposition 1.2.6 (Gaifman and Snir, 1982, Basic Fact 1.3)**.** *Suppose $\mathcal{L}$ is a language extending the language of (Peano-)arithmetic.*

p *is probabilistic and $\mathbb{N}$-additive iff there is some m a countably additive probability measure over* $\langle \mathsf{Mod}_{\mathcal{L}}^{\mathbb{N}}, \mathcal{F} \rangle$, *where $\mathcal{F}$ is the σ-field generated by* $\{\{\mathcal{M} \in \mathsf{Mod}_{\mathcal{L}}^{\mathbb{N}} \mid \mathcal{M} \models \varphi\} \mid \varphi \in \mathsf{Sent}_{\mathcal{L}}\}$, *such that*

$$\mathrm{p}(\varphi) = m\{\mathcal{M} \in \mathsf{Mod}_{\mathcal{L}}^{\mathbb{N}} \mid \mathcal{M} \models \varphi\}.$$

---

[12]If the extended language also has a natural number predicate, we will then say that p is $\mathbb{N}$-additive if

$$\mathrm{p}(\exists x \in N\ \varphi(x)) = \lim_{n} \mathrm{p}(\varphi(\overline{0}) \vee \ldots \vee \varphi(\overline{n})).$$

We will sometimes be considering *non-classical probabilities* where one may drop the requirement that the models of the language be classical models. In that case we might use an alternative axiomatisation. Those axioms are:

- $p(\top) = 1$,

- $p(\bot) = 0$,

- $p(\varphi \wedge \psi) + p(\varphi \vee \psi) = p(\varphi) + p(\psi)$,

- If $\varphi \vDash \psi$ then $p(\varphi) \leqslant p(\psi)$.

For the case where we consider logical consequence, $\vDash$, is as given in classical logic, this is equivalent to the standard axiomatisation.

1.2.2 Which interpretation

Now we've presented certain axioms for probability, but probability is interesting because it can be used to characterise something. There are many different applications of the probability notion. Three very important and influential ones are:

- Subjective probability, or the credences or degrees of belief of a (rational) agent. This measures the strength of an agent's beliefs.

- Objective chances. These identify some property of the world.

- Evidential probability, or logical probability. This measures the strength of an argument for a conclusion.

The application of probability that I am particularly interested in is subjective probability, or an agent's degrees of belief. This interpretation will be focused on throughout this thesis. The discussion in Part II is specific to this interpretation. However much of what we will be saying in Part I won't be specific to that interpretation and will apply to the other notions, particularly to the notion of objective chances.

If one is interested in subjective probability when developing a semantics, one will want the flexibility to talk about different agent's beliefs. We will therefore generally add the probability facts directly into the model to allow for this flexibility. We do this by basing the semantic construction on possible world structures. This is the sort of semantics that we develop in Chapters 3 and 4 and Section 5.3. In these semantics we will generalise theories of truth by adding in probability. These constructions don't then tell us anything more about truth.[13]

There is another aim that one might have when developing a semantics for probability: to develop a notion of probability that can tell us something additional about, for example, the liar-paradox, and how truth and self-reference act. This may involve measuring how paradoxical a sentence is. One might then work with an alternative interpretation or application of the probability notion. We might call this interpretation:

---

[13]As presented in Section 3.4.3 for probabilistic modal structures which have no contingent vocabulary, the semantics is then just the truth semantics with the probability being 1 if the sentence is true, 0 if false and $[0, 1]$ otherwise.

- *Semantic probability*

This would be closely related to a degrees of truth theory. It might, for example, say that the probability of the liar sentence, λ, is $^1/_2$.[14] A semantics capturing this use of probability is presented in Section 5.2, extending Leitgeb (2012) where the approximate idea is that the probability of a sentence is given by the proportion of the stages of the revision sequence in which the sentence is satisfied. Another semantics that would fall into this category would be to take the probability of a sentence to be determined by the "proportion of" the Kripke fixed points in which it is true. We will not develop or further mention such a semantics in this thesis.

1.3 Connection to the Liar paradox

The liar sentence is a sentence that says that it is not true.

In our formal framework, we will capture this by a sentence, called λ, for which

$$\lambda \leftrightarrow \neg \mathsf{T} \ulcorner \lambda \urcorner$$

is arithmetically derivable. As we have mentioned $\ulcorner \varphi \urcorner$ is a way for our language to refer to the sentence φ, which it does by referring to a natural number that is the code of the sentence. By the diagonal lemma such a sentence will exist in the language $\mathcal{L}_\mathsf{T}$, where T is added as a predicate to the base language $\mathcal{L}$ (see Section 1.6 for further details).

The liar sentence causes a paradox because it leads to contradictions under basic assumptions about truth. Tarski argued that the T-biconditionals, $\mathsf{T}\ulcorner\varphi\urcorner \leftrightarrow \varphi$, are constitutive of our notion of truth and it is these that lead to contradiction due to the instance for the liar sentence: $\lambda \leftrightarrow \mathsf{T}\ulcorner\lambda\urcorner$.

Proposition 1.3.1. *The principles* $\mathsf{T}\ulcorner\varphi\urcorner \leftrightarrow \varphi$ *for all sentences* $\varphi \in \mathsf{Sent}_{\mathcal{L}_\mathsf{T}}$ *lead to inconsistencies in classical logic.*

This paradox has led to a vast amount of research on the question of how the truth predicate should work.

Our probabilistic liar, π, is taken to be a sentence such that

$$\pi \leftrightarrow \neg \mathsf{P}_{\geqslant 1/2} \ulcorner \pi \urcorner$$

is arithmetically derivable. Again, by the diagonal lemma, this will exist in a language, $\mathcal{L}_{\mathsf{P}_{\geqslant r}}$ where we add predicates $\mathsf{P}_{\geqslant r}$ for some class of real numbers.[15]

There is just at first glance a syntactical similarity between λ and π. π doesn't lead to problems under such basic assumptions as λ does, but it does lead to conflicts between seemingly harmless principles, for example a principle of introspection, which we will present in the next section, or deference, which we will present in Section 1.7.1.

Throughout this thesis we will be using techniques developed for dealing with the liar paradox to help us deal with the probabilistic liar.

---

[14]And perhaps also: sentences that are grounded should receive probability 1 or 0 depending on whether they are true or false.

[15]See Definition 1.6.7.

1.4 The problem with introspection

There is a conflict between probabilism and introspection that has been discussed in Caie (2013); Christiano et al. (ms); Campbell-Moore (2015b). Although one might not want to support introspection, it seems to be a bad feature that it is contradictory in this way.

Theorem 1.4.1. $\mathrm{p} : \mathsf{Sent}_{\mathcal{L}_{\mathsf{P}_{\geqslant r}}} \to \mathbb{R}^{16}$ *cannot satisfy all:*

- p *is probabilistic over Peano Arithmetic,* PA*, in particular:*

    - *If* $\mathsf{PA} \vdash \varphi \leftrightarrow \psi$*, then* $\mathrm{p}(\varphi) = \mathrm{p}(\psi)$*,*
    - $\mathrm{p}(\varphi) = 1 \implies \mathrm{p}(\neg\varphi) = 0$*.*

- *Introspection:*

    - $\mathrm{p}(\varphi) \geqslant r \implies \mathrm{p}(\mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner) = 1$
    - $\mathrm{p}(\varphi) \ngeqslant r \implies \mathrm{p}(\neg\mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner) = 1$

Where φ and ψ may be any sentences of $\mathcal{L}_{\mathsf{P}_{\geqslant r}}$.
    Similarly, the following theory is inconsistent (over classical logic):

-
    - $\mathsf{Prov}_{\mathsf{PA}}(\ulcorner\varphi \leftrightarrow \psi\urcorner) \to (\mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner \leftrightarrow \mathsf{P}_{\geqslant r}\ulcorner\psi\urcorner)$
    - $\mathsf{P}_{=1}\ulcorner\varphi\urcorner \to \neg\mathsf{P}_{\geqslant 1/2}\ulcorner\neg\varphi\urcorner$

-
    - $\mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner \to \mathsf{P}_{=1}\ulcorner\mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner\urcorner$
    - $\neg\mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner \to \mathsf{P}_{=1}\ulcorner\neg\mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner\urcorner$

Proof. Consider π with
$$\mathsf{PA} \vdash \pi \leftrightarrow \neg\mathsf{P}_{\geqslant 1/2}\ulcorner\pi\urcorner.$$

$$\begin{aligned}
\mathrm{p}(\pi) \geqslant 1/2 &\implies \mathrm{p}(\mathsf{P}_{\geqslant 1/2}\ulcorner\pi\urcorner) = 1\\
&\implies \mathrm{p}(\neg\mathsf{P}_{\geqslant 1/2}\ulcorner\pi\urcorner) = 0\\
&\implies \mathrm{p}(\pi) = 0\\
\mathrm{p}(\pi) \ngeqslant 1/2 &\implies \mathrm{p}(\neg\mathsf{P}_{\geqslant 1/2}\ulcorner\pi\urcorner) = 1\\
&\implies \mathrm{p}(\pi) = 1
\end{aligned}$$

So any assignment of probability to π leads to a contradiction. The second result holds because all this reasoning can be formalised by this theory. □

We will discuss this later in Sections 2.4.2, 3.4.1 and 5.4.2. In particular we will argue in Section 3.4.1 that this formalisation of the idea of introspection is wrong and it should instead be formulated as:

$$\mathsf{T}\ulcorner\mathsf{P}_{\geqslant}(\varphi, \ulcorner r\urcorner)\urcorner \to \mathsf{P}_{=}(\ulcorner\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner, \ulcorner 1\urcorner)$$
$$\text{and } \mathsf{T}\ulcorner\neg\mathsf{P}_{\geqslant}(\varphi, \ulcorner r\urcorner)\urcorner \to \mathsf{P}_{=}(\ulcorner\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner, \ulcorner 1\urcorner)$$

where one adopts a consistent (therefore non-transparent) theory for the truth predicate.[17] So we suggest rejecting the principle of introspection, but not by rejecting the motivation behind it but instead by rejecting its formulation.

---

[16]$\mathsf{Sent}_{\mathcal{L}_{\mathsf{P}_{\geqslant r}}}$ means the sentences of the language $\mathcal{L}_{\mathsf{P}_{\geqslant r}}$. See Section 1.6.3.
[17]In such a theory, $\mathsf{T}\ulcorner\neg\mathsf{P}_{\geqslant 1/2}\ulcorner\pi\urcorner\urcorner \leftrightarrow \neg\mathsf{P}_{\geqslant 1/2}\ulcorner\pi\urcorner$ cannot hold.

A very similar result was used in Caie (2013) to argue that probabilism should be rejected. In fact we do argue in Chapters 3 and 4 that probabilism should be rejected but not in the way that Caie suggests, instead in a more cautious way. We will reject probabilism by instead saying that an agent should be *non-classically* probabilistically coherent.

In Egan and Elga (2005), closely related contradictions were taken to show that in situations like *Promotion* an agent should not believe the biconditional *Promotion* $\leftrightarrow \neg \mathbb{P}_{\geqslant 1/2} Promotion$, i.e. she cannot take herself to be an anti-expert. In fact the contradiction does not require that the agent believe *Promotion* $\leftrightarrow \neg \mathbb{P}_{\geqslant 1/2} Promotion$, it just requires that she assigns the same degree of belief to *Promotion* as she does to $\neg \mathbb{P}_{\geqslant 1/2} Promotion$. As discussed in Section 1.1.1, the analogous response isn't available in the case of π. For the case of π, the biconditional is arithmetically derivable so an agent should believe it (so long as they are probabilistically certain about Robinson arithmetic).[18] So we have to account for a sentence like π where an agent cannot satisfy this introspection principle and perhaps then we can apply such considerations to cases like *Promotion* and allow an agent to learn they are an anti-expert.

1.5 Questions to answer and a broad overview

A very important question to answer when we have self-referential probabilities is:

> How can one develop a formal semantics for this language?

A semantics tells one when sentences are true and false. This is important to do to better understand the languages that we are working with. This will also help us see which principles are inconsistent in the self-referential framework and how to modify principles that are inconsistent to allow them to be consistent. When developing a semantics one should also look to develop axiomatic theories for the language. We consider these questions in Part I.

In Part I there are four chapters. The first provides an introduction to the challenge of providing a semantics and the strategy we will use in the other chapters. This is to apply theories of truth to possible world structures, in the form of probabilistic modal structures. These structures immediately allow for a definition of a semantics in the operator language but we will show that the analogous definition cannot be applied in the predicate setting as there are often no *Prob-PW-models*. Chapters 3 and 4 develop Kripke-style theories of probability which require one to move to non-standard probabilities but allows one to obtain models which are in a certain sense stable. Chapter 5 develops a revision theory of probability where we retain classical probabilities but result in a transfinite sequence of models. A more detailed overview of these chapters can be found in Section 2.6.

Another important question is:

> What rationality constrains are there on an agent once such expressive languages are considered?

---

[18]Furthermore, if we use the strong diagonal lemma we can show that there is some term t of L_{PA} such that $\mathsf{PA} \vdash t = \ulcorner \neg \mathsf{P}_{\geqslant 1/2} t \urcorner$, and therefore if we just assume that arithmetic is true, then we have $\mathsf{P}_{\geqslant r} t \leftrightarrow \mathsf{P}_{\geqslant r} \ulcorner \neg \mathsf{P}_{\geqslant 1/2} t \urcorner$.

and, relatedly,

> To what degree should a rational agent believe a sentence that says something about her own degrees of belief?

These questions focus on the subjective interpretation of probability. In Part II, we consider these questions. For such self-referential sentences, a choice of the agent's credences will affect which worlds are possible. Caie (2013) has argued that the accuracy and Dutch book arguments should be modified because the agent should only care about her inaccuracy or payoffs in the world which could be actual if she adopted the considered credences. We consider the accuracy argument in Chapter 7 and the Dutch book argument in Chapter 8. Much of Chapter 8 is taken to determining a criterion saying that an agent should try to minimize his overall guaranteed losses, assuming he bets with his credences. Both these accuracy and Dutch book criteria mean that an agent is rationally required to be probabilistically incoherent (and not be representable in our suggested semantics), have negative credences and to fail to assign the same credence to logically equivalent sentences. We will also show that this accuracy criterion depends on how inaccuracy is measured and that the accuracy criterion differs from the Dutch book criterion.

This also has connections to the first question if we want our semantics to be able to represent rational agents. If rationality considerations say that an agent should be one way, our semantics should allow the probability notion to be that way. In fact we will reject Caie's suggested modifications. In the final section of Chapter 7, Section 7.3, we reconsider the accuracy criterion and instead suggest that the agent should consider how accurate her credences are from the perspective of her current credences. We will also consider how to generalise this version of the accuracy criteria and present ideas suggesting that it connects to the semantics developed in Part I suggesting that our semantics can represent rational agents. In the final section of Chapter 8, Section 8.6, we suggest that this is a case where an agent should not bet with his credences.

Before starting on Part I, we will first present some technical preliminaries and briefly mention conditional probabilities, which won't generally be considered.

1.6 Technical preliminaries

Throughout the thesis we will be using formal languages that can express self-referential sentences such as π. In analysing these sentences we will be using much of the terminology, as well as techniques and tools developed for the liar paradox. For example we will generally be following the notation from Halbach (2014).

1.6.1 Arithmetisation

Setup 1. *We will consider first order languages to have the logical constants $=$, $\vee$, $\neg$ and $\exists$. The other connectives, $\wedge$, $\forall$ and $\rightarrow$, will be taken as defined as usual, e.g. $\varphi \wedge \psi := \neg(\neg\varphi \vee \neg\psi)$.*

For an introduction to first order logic, see any introductory logic textbook.

We will often work with formal languages $\mathcal{L}$ that extend the language of Peano Arithmetic. More generally we could allow $\mathcal{L}$ to just be able to code arithmetic, but for simplicity we will simply assume that we have Peano arithmetic directly available in our language.

Definition 1.6.1. L_{PA} is a first-order language with identity with the constant 0, standing for zero, a one-place function symbol S standing for the successor function taking n to $n+1$, and two pace function symbols $+$ and $\times$ for addition and multiplication. We shall also assume we have a number of special additional (non-contingent) predicates and function symbols available in L_{PA} as specified in Definition 1.6.2.

A language $\mathcal{L}$ extends L_{PA} if it adds additional vocabulary, which we will sometimes refer to as contingent or empirical vocabulary. Sometimes we will also allow that these languages contain a natural number predicate N, but if so that will be explicitly stated.

The theory of Peano arithmetic, PA is given by the defining equations for zero, successor, addition and multiplication as axioms. PA should also contain the defining axioms for the additional predicates and function symbols added to L_{PA} (in Definition 1.6.2). Finally, PA contains the induction axioms, which are all sentences of the form

$$\varphi(0) \wedge \forall x(\varphi(x) \rightarrow \varphi(\mathsf{S}(x))) \rightarrow \forall x \varphi(x).$$

For a language $\mathcal{L}$ extending L_{PA} the theory $\mathsf{PA}_{\mathcal{L}}$ extends the theory of PA as the induction principle includes an instance for each φ a formula of $\mathcal{L}$. If the extended language contains a natural number predicate, N, then all the quantifiers in the theory of PA are restricted to this predicate, for example, the induction principle is then:

$$\varphi(0) \wedge \forall x(N(x) \rightarrow (\varphi(x) \rightarrow \varphi(\mathsf{S}(x)))) \rightarrow \forall x(N(x) \rightarrow \varphi(x)).$$

We will often drop the explicit reference to $\mathcal{L}$, and when PA is talked about in the context of an extended language $\mathsf{PA}_{\mathcal{L}}$ will be meant.

$\mathbb{N}$ refers to the standard model of PA. This contains just the numbers 0, 1, 2 etc and interprets the vocabulary as intended.

Robinson arithmetic, or Q, is the theory in L_{PA} which replaces the induction scheme by[19]

$$\forall x(x \neq 0 \rightarrow \exists z(z + x = y)).$$

We will use arithmetic to provide us with objects with which to refer to sentences of the language and to assign probabilities to these sentences. An assumption that we will have to make to be able to do this is that the language that we work with is countable and that their syntax is recursive. This is an assumption that we may wish to drop but we will then need an alternative syntax-theory. To keep things simple we will therefore assume that every language we discuss is countable and has recursive syntax and we will not generally explicitly note this restriction.

Definition 1.6.2. Take some language $\mathcal{L}$ (that is countable with recursive syntax).

---

[19]Again, if $\mathcal{L}$ contains a natural number predicate then all quantifiers in Q will be restricted to N.

We denote the set of sentences of $\mathcal{L}$ by $\mathsf{Sent}_\mathcal{L}$ and the set of formulas of $\mathcal{L}$ by $\mathsf{Form}_\mathcal{L}$. In Section 1.6.3 we will introduce languages like $\mathcal{L}_{\mathsf{P,T}}$ extending some base language, $\mathcal{L}$ with extra predicates etc to represent probability or truth. For these languages, if $\mathcal{L}$ is clear from context we will drop the explicit reference to it, for example writing $\mathsf{Sent}_{\mathsf{P,T}}$ instead of to $\mathsf{Sent}_{\mathcal{L}_{\mathsf{P,T}}}$, or $\mathsf{Sent}_{\mathsf{P}_\geqslant}$ instead of $\mathsf{Sent}_{\mathcal{L}_{\mathsf{P}_\geqslant}}$, and similarly for Form.

We assume some coding $\#$ of expressions of $\mathcal{L}$ to $\mathbb{N}$ that is recursive and one-to-one, so $\#\varphi$ stands for the number which codes φ. Details of coding can be found in textbooks that include an account of Gödel's incompleteness theorems. For $\varphi \in \mathsf{Form}_\mathcal{L}$, we let $\ulcorner\varphi\urcorner$ denote the *numeral*[20] corresponding to $\#\varphi$.

We shall assume that we have predicates in L_{PA} (strongly) representing $\mathsf{Sent}_\mathcal{L}$ and $\mathsf{Form}_\mathcal{L}$, so for example $\mathsf{Sent}_\mathcal{L}(\overline{n})$ is a theorem of PA iff $n = \#\varphi$ for some $\varphi \in \mathsf{Sent}_\mathcal{L}$.[21]

If $\triangleright$ is a syntactic operation we will denote its corresponding operation on natural numbers by and assume that we have this as a function symbol $\dot{\triangleright}$ in L_{PA} representing it. For example $\ulcorner\varphi\urcorner \dot{\vee} \ulcorner\psi\urcorner = \ulcorner\varphi \vee \psi\urcorner$ is a theorem of PA, and so is $\dot{\neg}\ulcorner\varphi\urcorner = \ulcorner\neg\varphi\urcorner$. Also, e.g. if $\mathsf{P}_{\geqslant r}$ is a predicate in $\mathcal{L}$, then $\dot{\mathsf{P}}_{\geqslant r}\ulcorner\varphi\urcorner = \ulcorner\mathsf{P}_{\geqslant r}\varphi\urcorner$ is a theorem of PA.

For a term t of L_{PA} we denote the interpretation of the term t in the standard model of arithmetic, $\mathbb{N}$, by $t^\mathbb{N}$. For example $(\mathsf{S}\overline{n})^\mathbb{N} = n + 1$. We will also represent the interpretation function by $\circ$, but this is understood not to be a function symbol in our language. We therefore have that for any term t of L_{PA}, $\ulcorner t\urcorner^\circ = \ulcorner\overline{t^\mathbb{N}}\urcorner$ is a true, non-atomic formula of L_{PA}.

The substitution function will be represented by $x(y/z)$, so $\ulcorner\varphi\urcorner(\ulcorner t\urcorner/\ulcorner v\urcorner) = \ulcorner\varphi(t/v)\urcorner$ is a theorem of PA, where $\varphi(t/v)$ denotes the formula φ with all instances of the variable v replaced by the term t.

Using this notation and setup we present the Diagonal lemma which informally says that there must be sentences which talk about themselves.

Theorem 1.6.3 (Diagonal Lemma)**.** *Let $\mathcal{L}$ be a countable, recursive language extending L_{PA}. Let $\varphi(v)$ be a formula of $\mathcal{L}$ with v the only free variable in φ. Then there is a sentence δ such that:*

$$\mathsf{PA} \vdash \delta \leftrightarrow \varphi(\ulcorner\delta\urcorner)$$

Proof. Let sub be such that, if $\psi(v)$ has only v free,

$$\mathsf{PA} \vdash \mathsf{sub}(\ulcorner\psi(v)\urcorner, \overline{n}) = \ulcorner\psi(\overline{n})\urcorner.$$

Suppose φ has only v free. Then let $\delta := \varphi(\mathsf{sub}(\ulcorner\varphi(\mathsf{sub}(v,v))\urcorner, \ulcorner\varphi(\mathsf{sub}(v,v))\urcorner))$.

Now

$$
\begin{array}{rl}
\delta \leftrightarrow & \varphi(\mathsf{sub}(\ulcorner\varphi(\mathsf{sub}(v,v))\urcorner, \ulcorner\varphi(\mathsf{sub}(v,v))\urcorner)) \\
\mathsf{PA} \vdash \quad \leftrightarrow & \varphi(\ulcorner\varphi(\mathsf{sub}(\ulcorner\varphi(\mathsf{sub}(v,v))\urcorner, \ulcorner\varphi(\mathsf{sub}(v,v))\urcorner))\urcorner) \qquad\qquad \square \\
\leftrightarrow & \varphi(\ulcorner\delta\urcorner)
\end{array}
$$

---

[20] The numeral of n is denoted $\overline{n}$ and it corresponds to the expression $\overbrace{\mathsf{S}(\ldots\mathsf{S}}^{n}(0)\ldots)$.

[21] And otherwise $\neg\mathsf{Sent}_\mathcal{L}(\overline{n})$ is a theorem of PA. I.e. it is *strongly* represented. This will hold for all our representations here, but we will not explicitly mention that fact.

Corollary 1.6.4. *Suppose* $\mathcal{L}$ *extends* L_{PA} *and contains a predicate* $\mathsf{P}_{\geqslant 1/2}$. *Then there must be a* probabilistic liar, π, *where*

$$\mathsf{PA} \vdash \pi \leftrightarrow \neg \mathsf{P}_{\geqslant 1/2}\ulcorner \pi \urcorner$$

And a probabilistic truthteller η, *where*

$$\mathsf{PA} \vdash \eta \leftrightarrow \mathsf{P}_{\geqslant 1/2}\ulcorner \eta \urcorner$$

If the language has a predicate $\mathsf{P}_{=1}$, *it will contain* γ *where*

$$\mathsf{PA} \vdash \gamma \leftrightarrow \neg \forall n \in N \overbrace{\mathsf{P}_{=1}\ulcorner \mathsf{P}_{=1}\ulcorner \ldots \mathsf{P}_{=1}\ulcorner \gamma \urcorner \urcorner \urcorner}^{n+1}.$$

If $\mathcal{L}$ *extends* L_{PA} *and contains a predicate* T, *there is a* liar, λ, *where*

$$\mathsf{PA} \vdash \lambda \leftrightarrow \neg \mathsf{T}\ulcorner \lambda \urcorner.$$

And a truthteller, τ, *where*

$$\mathsf{PA} \vdash \tau \leftrightarrow \mathsf{T}\ulcorner \tau \urcorner.$$

1.6.2 Reals

Since we are dealing with probabilities, which take values in $\mathbb{R}$, we will sometimes work with languages that can refer to such probability values, typically by assuming we have a language that extends L_{ROCF}. We will then also assume that we have in the background the theory ROCF.

Definition 1.6.5. Let the language of real ordered closed fields, L_{ROCF}, be a first order langauge with $+$, $-$, $\times$ (sometimes written as $\cdot$), 0, 1 and $<$.

Let $\mathbb{R}$ denote the intended model of this language. This has the real numbers as the domain, and interprets the vocabulary as intended.

The theory ROCF is:

- Field axioms[22]

- Order axioms:

    - $\forall x, y, z(x < y \rightarrow x + z < y + z)$
    - $\forall x, y, z((x < y \wedge 0 < z) \rightarrow x \cdot z < y \cdot z)$

- Roots of polynomials[23]

    - $\forall x(x > 0 \rightarrow \exists y(y^2 = x))$,
    - for any n odd:

    $$\forall x_0, \ldots x_n(x_n \neq 0 \rightarrow \exists y(x_0 + x_1 \cdot y + x_2 \cdot y^2 + \ldots + x_n \cdot y^n = 0))$$

---

[22]See any introductory mathematical logic textbook (For example, Margaris, 1990, p. 115, example 7. The theory ROCF can be found in example 11).

[23]y^n is shorthand for $\overbrace{y \cdot \ldots \cdot y}^{n}$.

The rational numbers are all real numbers of the form $^m/_n$ for $m, n \in \mathbb{N}$. The collection of all the rational numbers is denoted $\mathbb{Q}$. Being equal to the rational number r is definable in ROCF, so we will assume that L_{ROCF} has terms denoting each rational number, $\bar{r}$.

Let $L_{\mathsf{PA,ROCF}}$ be the join of the two languages. We will then assume that we have predicates N and R, where PA is restricted to N and ROCF is restricted to R.[24]

In the language $\mathcal{L}_{\mathsf{P}_{\geqslant}}$, which we will introduce in Section 1.6.3, we will work over arithmetic but refer to probability values by using a coding that also codes the rational numbers. This will work as follows.

Definition 1.6.6. In the context of the language $\mathcal{L}_{\mathsf{P}_{\geqslant}}$ we will assume that our coding $\#$ codes not only the expressions of the language, but also the rational numbers in a recursive one-to-one manner. So for r a rational number, we have $\#r$ is a natural number which is the code of r. As for codings of the language, we will use $\ulcorner r \urcorner$ to denote the numeral corresponding to $\#r$.

We use $\mathsf{rat}(n)$ to denote the rational number whose code is n. So $\mathsf{rat}(\#r) = r$.

We shall use Rat to denote the set of codes of rational numbers. We also assume that we have this as a predicate available in L_{PA}. So $\mathsf{Rat}(\bar{n})$ is a theorem of PA iff there is some rational number r with $n = \#r$.[25]

We shall represent operations on the rationals by the subdot notation as we did for syntactic operations, for example we have a function symbol $\dotplus$ where $\mathsf{PA} \vdash \ulcorner r \urcorner \dotplus \ulcorner q \urcorner = \ulcorner r + q \urcorner$. We shall use $\prec$ to represent the ordering on the rational numbers, so $\#r \prec \#q \iff r < q$, and assume that $\prec$ is available in our language. So $\ulcorner r \urcorner \prec \ulcorner q \urcorner$ is a theorem of PA iff $r < q$.

1.6.3 The languages we consider

We now present the formal languages that will be considered throughout this thesis.

Expressively rich probability languages

There are a few different options for how one can formulate probability within a first order language. The first is to do something analogous to as is often done in the operator case and add predicates $\mathsf{P}_{\geqslant r}$ for each rational number, r. The second is to consider adding a binary predicate $\mathsf{P}_{\geqslant}$ and having rational numbers, or representations thereof, also in the base language. The third is to consider P as a function symbol.

We start with the first, and simplest, of these languages.

Definition 1.6.7. Let $\mathcal{L}$ be some (countable[26]) language extending L_{PA}.

We will let $\mathcal{L}_{\mathsf{P}_{\geqslant r}}$ denote the extension of $\mathcal{L}$ by the countably many unary predicates $\mathsf{P}_{\geqslant r}$ and $\mathsf{P}_{>r}$ for each rational number r with $0 \leqslant r \leqslant 1$.

---

[24]It is not important for our considerations if we have two copies of the common vocabulary or if the single vocabulary of ROCF is also used to also interpret PA when restricted to the predicate N. For simplicity we will assume we have distinct languages for the two domains.

[25]And $\mathsf{PA} \vdash \neg\mathsf{Rat}(\bar{n})$ if not.

[26]Remember we have to assume the languages are countable for the Gödel numbering to work. From now on we will not generally explicitly mention this in the specifications of our languages but it will always be implicitly assumed.

We can then take the other predicates like $\mathsf{P}_{\leqslant r}$ to be defined:

Definition 1.6.8. Define:

- $\mathsf{P}_{\leqslant r}t := \mathsf{P}_{\geqslant 1-r}\ulcorner\neg t$

- $\mathsf{P}_{<r}t := \mathsf{P}_{>1-r}\ulcorner\neg t$

- $\mathsf{P}_{=r}t := \mathsf{P}_{\geqslant r}t \wedge \mathsf{P}_{\leqslant r}t$

This style of definition of these derivative notions follows Heifetz and Mongin (2001). One could also just take $\mathsf{P}_{\geqslant r}$ as primitive and take $\mathsf{P}_{>r}$ as defined by $\mathsf{P}_{>r}t := \mathsf{Sent}_{\mathsf{P}_{\geqslant r}}(t) \wedge \neg \mathsf{P}_{\leqslant r}t$. If we are dealing with an interpretation of probability that is classical then this will be equivalent, however, the advantage of also having the primitive predicate $\mathsf{P}_{>r}$ is that one can also use this language in the non-classical settings where one can have for some φ and r with neither $\mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner$ nor $\mathsf{P}_{\leqslant r}\ulcorner\varphi\urcorner$ being satisfied. This is, for example, the kinds of probabilities which we study in Chapter 3. In that setup, $\neg\mathsf{P}_{\leqslant r}\ulcorner\varphi\urcorner$ will no longer capture the intended meaning of $\mathsf{P}_{>r}\ulcorner\varphi\urcorner$.

The predicate setting allows for greater expressive power, as discussed in Section 1.1.1, as one can for example have a sentence

$$\exists x(\mathsf{P}_{>0}x \wedge \mathsf{P}_{<1}x)$$

which would not be expressible in the operator framework. Note that this does allow us to get higher order probabilities because the Gödel numbering will code all sentences of the expanded language $\mathcal{L}_{\mathsf{P}_{\geqslant r}}$. We will also obtain self-referential probabilities in this language via the Diagonal Lemma (Theorem 1.6.3). So for example we have a sentence π where:

$$\mathsf{PA} \vdash \pi \leftrightarrow \neg\mathsf{P}_{\geqslant 1/2}\ulcorner\pi\urcorner.$$

We will always take a certain amount of arithmetic as assumed when we are considering $\mathcal{L}_{\mathsf{P}_{\geqslant r}}$ because to use $\mathcal{L}_{\mathsf{P}_{\geqslant r}}$ to express facts about probability we need the objects to which probabilities are attached to be represented in this language, which we do by means of Gödel coding as in Definition 1.6.2.

This is the language that we will generally work with, throughout this thesis, when we are interested in *negative results* saying that adding certain principles, like the introspection principle (Section 1.4) or a deference principle (Section 1.7), leads to inconsistency when such self-referential sentences are around. This is because the language is the simplest and expressively weakest of the languages that we consider which can express such sentences, so this leads to stronger inconsistency results: even in such a restrictive language these principles can lead to contradictions.

One important step from the operator to the predicate setting is this quantificational ability. However this langauge $\mathcal{L}_{\mathsf{P}_{\geqslant r}}$ still has limited quantificational ability because it cannot quantify over the probability values. To allow for such quantification we therefore need the probability values, or representations thereof, to be present in the domain. The next language we present allows for quantification over the probability values and it represents the probability values by just assuming that we have a domain containing the real numbers. This language extends $\mathcal{L}_{\mathsf{P}_{\geqslant r}}$.

Definition 1.6.9. Let $\mathcal{L}$ be some (countable) language extending $L_{\mathsf{PA,ROCF}}$.

Let $\mathcal{L}_{\mathsf{P}_\geqslant}^{\mathbb{R}}$ denote the extension of $\mathcal{L}$ by a binary predicate $\mathsf{P}_\geqslant(\cdot,\cdot)$. Here $\mathsf{P}_\geqslant(\ulcorner\varphi\urcorner, \overline{r})$ will stand for "the probability of φ is $\geqslant r$".

Note that in this language we did not add $\mathsf{P}_>$ as primitive, this is because the quantificational ability present in $\mathcal{L}_{\mathsf{P}_\geqslant}$ will allow us to define $\mathsf{P}_>$ by:

$$\mathsf{P}_>(s,t) \leftrightarrow \exists x > t\, \mathsf{P}_\geqslant(s,x)$$

then the usual instances of this defining equality are

$$\mathsf{P}_>(\ulcorner\varphi\urcorner, \overline{r}) \leftrightarrow \exists x > \overline{r}\, \mathsf{P}_\geqslant(\ulcorner\varphi\urcorner, \overline{r})$$

Intended models of $\mathcal{L}_{\mathsf{P}_\geqslant}^{\mathbb{R}}$ will always have domain interpreting R with $\mathbb{R}$ and N with $\mathbb{N}$.

In fact we will not be studying $\mathcal{L}_{\mathsf{P}_\geqslant}^{\mathbb{R}}$ in this thesis, instead we will consider a closely related language where instead of having real numbers as themselves in the domain we work with natural number codes for real numbers. Such a language will be used in Chapter 3. This allows us to just work with a theory of natural numbers. Of course there are only countably many natural numbers and uncountably many real numbers that we wish to represent, so we will instead just code up the rational numbers.

Definition 1.6.10. Let $\mathcal{L}$ be some (countable) language extending L_{PA}.

Assume we have a coding $\#$ that, as well as coding natural numbers, codes the rational numbers into the naturals in a recursive way.

Let $\mathcal{L}_{\mathsf{P}_\geqslant}$ denote the extension of $\mathcal{L}$ by a binary predicate $\mathsf{P}_\geqslant(\cdot,\cdot)$. Here $\mathsf{P}_\geqslant(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)$ will stand for "the probability of φ is $\geqslant r$".

We can then again take the other facts about probability as defined, now defining

$$\mathsf{P}_>(s,t) \leftrightarrow \exists x \succ t\, \mathsf{P}_\geqslant(s,x)$$

where the usual instances are

$$\mathsf{P}_>(\ulcorner\varphi\urcorner, \ulcorner r\urcorner) \leftrightarrow \exists x \succ \ulcorner r\urcorner\, \mathsf{P}_\geqslant(\ulcorner\varphi\urcorner, \ulcorner r\urcorner).$$

Here we only have codes for the *rational numbers* in our domain, but we may still have models that are interpreted as assigning irrational probability values to sentences. For example there may be a sentence φ where for each $r < \sqrt{2}/2$, $\mathsf{P}_\geqslant(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)$ is satisfied, and for each $r > \sqrt{2}/2$, $\mathsf{P}_\leqslant(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)$ is satisfied. The restriction is just that $\mathsf{P}_=\left(\ulcorner\varphi\urcorner, \ulcorner\sqrt{2}/2\urcorner\right)$ is not a sentence of the language.

Although the language is closely related to $\mathcal{L}_{\mathsf{P}_\geqslant}^{\mathbb{R}}$, it is different. For example, in intended models of $\mathcal{L}_{\mathsf{P}_\geqslant}^{\mathbb{R}}$ we will have[27]

$$\forall x(\mathsf{Sent}_\mathcal{L}(x) \rightarrow \exists y \mathsf{P}_=(x,y))$$

as true, whereas this will *not* be true in all intended models of $\mathcal{L}_{\mathsf{P}_\geqslant}$ because we may want models which we can interpret as assigning irrational numbers

---

[27]In this example we restrict attention to the probability of sentences in $\mathcal{L}$. If we take models of $\mathcal{L}_{\mathsf{P}_\geqslant}^{\mathbb{R}}$ where the probability notion is taken to be a classical probability then we will have this for all sentences of $\mathcal{L}_{\mathsf{P}_\geqslant}^{\mathbb{R}}$, however we do not do this so we can also apply these considerations in Chapter 3 where we consider more generalised probabilities.

to sentences though there is no variable assignment which can witness that probability value as there is nothing in the domain standing for that irrational number.

As we have already mentioned, we can use the above languages when we wish to have a theory of probability for a generalised notion of probability where one may assign ranges of probability values to single sentences. However, if that is not of interest and one wishes to enforce from the begging that sentences assign single point valued probability values, then one could formalise P as a function symbol.

Definition 1.6.11. Let $\mathcal{L}$ be some (countable) language extending $L_{\mathsf{PA,ROCF}}$.

Let $\mathcal{L}_{\mathsf{P}}$ denote the extension of $\mathcal{L}$ by a function symbol P.

Here $\mathsf{P}\ulcorner\varphi\urcorner = \bar{r}$ will stand for "the probability of φ is equal to r".

In this language one can naturally write relationships between probability values of different sentences, for example:

$$\mathsf{P}\ulcorner\varphi\wedge\psi\urcorner + \mathsf{P}\ulcorner\varphi\wedge\neg\psi\urcorner = \mathsf{P}\ulcorner\varphi\urcorner$$

or define

$$\mathsf{P}(\ulcorner\varphi\urcorner \,|\, \ulcorner\psi\urcorner) := \frac{\mathsf{P}\ulcorner\varphi\wedge\psi\urcorner}{\mathsf{P}\ulcorner\psi\urcorner}$$

where the right hand side is now expressible in the language. This language is basically equivalent to $\mathcal{L}_{\mathsf{P}_{\geqslant}}^{\mathbb{R}}$ except it builds in this assumption that sentences receive point valued probabilities. Suppose we have a theory in $\mathcal{L}_{\mathsf{P}_{\geqslant}}^{\mathbb{R}}$ with a theorem

$$\forall x(\mathsf{Sent}_{\mathsf{P}_{\geqslant}\mathbb{R}}(x) \rightarrow \exists! y \mathsf{P}_{=}(x, y)),$$

then one can also express, e.g.

$$\mathsf{P}\ulcorner\varphi\wedge\psi\urcorner + \mathsf{P}\ulcorner\varphi\wedge\neg\psi\urcorner = \mathsf{P}\ulcorner\varphi\urcorner$$

in a natural way by

$$\exists x, y, z(\mathsf{P}_{=}(\ulcorner\varphi\wedge\psi\urcorner, x) \wedge \mathsf{P}_{=}(\ulcorner\varphi\wedge\neg\psi\urcorner, y) \wedge \mathsf{P}_{=}(\ulcorner\varphi\urcorner, z) \wedge x + y = z).$$

When we want to make the assumption that sentences do receive point-valued probablity values this is the most natural language to work with, so it is what we use in Chapter 5. It is also used in Chapter 4 where probabilities are interpreted as sets of probability functions, but there we work with a non-classical semantics, so for some sentences none of the $\mathsf{P}\ulcorner\varphi\urcorner = \bar{r}$ will be true.

Languages with probability operators

In the existing literature on logics for probability, probability *operators* have been worked with. Such languages are restrictive and allow for higher order probabilities but not self-referential probabilities. In this thesis we will sometimes discuss the connection between the languages that we work with and the operator languages so we here present these operator languages.

There are a number of different operator languages that have been considered in the literature. (Aumann, 1999; Fagin et al., 1990; Ognjanović and Rašković, 1996; Bacchus, 1990; Heifetz and Mongin, 2001). We will focus mostly a language as set up in Heifetz and Mongin (2001). This is analogous to $\mathcal{L}_{\mathsf{P}_{\geqslant r}}$.

Definition 1.6.12. Consider the extension of a (possibly propositional) language $\mathcal{L}$ to $\mathcal{L}_{\mathbb{P}_{\geqslant r}}$ given by:

- If φ is a sentence of $\mathcal{L}$ then it is a sentence of $\mathcal{L}_{\mathbb{P}_{\geqslant r}}$,

- If φ and ψ are sentences of $\mathcal{L}_{\mathbb{P}_{\geqslant r}}$ then so are:

  - $\varphi \vee \psi$,
  - $\neg\varphi$,
  - $\mathbb{P}_{\geqslant r}\varphi$, for any rational number r, $0 \leqslant r \leqslant 1$,
  - $\mathbb{P}_{>r}\varphi$, for any rational number r, $0 \leqslant r \leqslant 1$.

Although we here have presented a language just with $\mathbb{P}_{\geqslant r}$ we can consider some other facts about the probability as defined. In this we follow Heifetz and Mongin (2001). These parallel the definitions in the case for $\mathcal{L}_{\mathsf{P}_{\geqslant r}}$.

Definition 1.6.13. Define:

- $\mathbb{P}_{\leqslant r}\varphi := \mathbb{P}_{\geqslant 1-r}\neg\varphi$

- $\mathbb{P}_{=r}\varphi := \mathbb{P}_{\geqslant r}\varphi \wedge \mathbb{P}_{\leqslant r}\varphi$

- $\mathbb{P}_{<r}\varphi := \mathbb{P}_{>1-r}\neg\varphi$

We will also, in Section 3.3.2, use a more expressive operator language that corresponds to $\mathcal{L}_{\mathsf{P}_{\geqslant}}$.

Definition 1.6.14. Let $\mathcal{L}$ be a language extending L_{PA}.

Consider the extension of a language $\mathcal{L}$ to $\mathcal{L}_{\mathbb{P}_{\geqslant}}$ given by:

- If φ is a formula of $\mathcal{L}$ then it is a formula of $\mathcal{L}_{\mathbb{P}_{\geqslant r}}$,

- If φ and ψ are formulas of $\mathcal{L}_{\mathbb{P}_{\geqslant r}}$ then so are:

  - $\varphi \vee \psi$,
  - $\neg\varphi$,
  - $\exists x\varphi(x)$,
  - $\mathbb{P}_{\geqslant}(\varphi, s)$, for any rational number r, $0 \leqslant r \leqslant 1$, and s a term of $\mathcal{L}$.

Languages with truth

Just as for probability, truth can be represented in a formal language either as an operator or in a first-order way. For probability there were many options for how probability could be represented in a first-order language, but for truth there is only one natural way: to represent it using a unary predicate T.

Definition 1.6.15. Let $\mathcal{L}_{\mathsf{T}}$ denote the extension of $\mathcal{L}$ by adding a unary predicate T, then $\mathsf{T}\ulcorner\varphi\urcorner$ will stand for "φ is true".

The language with an operator for truth can be defined as usual operator languages are defined:

Definition 1.6.16. For a language $\mathcal{L}$, let $\mathcal{L}_{\mathbb{T}}$ denote the extension of $\mathcal{L}$ with an operator $\mathbb{T}$, defined by:

- If φ is a sentence of $\mathcal{L}$ then it is a sentence of $\mathcal{L}_\mathbb{T}$,

- If φ and ψ are sentences of $\mathcal{L}_\mathbb{T}$ then so are:

  - $\varphi \vee \psi$,
  - $\neg\varphi$,
  - $\mathbb{T}\varphi$.

We will typically work with languages that can not only express truth but which can express both truth and probability.

Languages with truth and probability

If we want to have both truth and probability in our language we need to ensure that these notions can interact, so we can say things like:

"The probability of "$0 = 0$" is 1" is true

This will be a particular issue when we work with a language with both probability and truth as operators. If we just took an operator language, e.g. $\mathcal{L}_{\mathbb{P}_{\geqslant r}}$, and then added another operator to it, for example adding a truth operator in the manner according to Definition 1.6.16, then certain interactions of these notions would not be expressible, for example

$$\mathbb{P}_{\geqslant 1/2}(\mathbb{T}(0 = 0))$$

would not be a sentence according to the syntax rules. So instead if we want a language with multiple operators we will need to define the language simultaneously. For example:

Definition 1.6.17. The syntax of $\mathcal{L}_{\mathbb{P}_{\geqslant r},\mathbb{T}}$ is defined as:

- If φ is a sentence of $\mathcal{L}$ then it is a sentence of $\mathcal{L}_{\mathbb{P}_{\geqslant r},\mathbb{T}}$,

- If φ and ψ are sentences of the $\mathcal{L}_{\mathbb{P}_{\geqslant r},\mathbb{T}}$ then so are:

  - $\varphi \vee \psi$,
  - $\neg\varphi$,
  - $\mathbb{P}_{\geqslant r}\varphi$, for any rational number r, $0 \leqslant r \leqslant 1$,
  - $\mathbb{P}_{>r}\varphi$, for any rational number r, $0 \leqslant r \leqslant 1$,
  - $\mathbb{T}\varphi$.

When we have a language with both probability and truth represented as predicates these considerations will automatically be taken care of by the assumption that we are using a coding which codes all the sentences of the language at stake. For example we can just define the syntax of $\mathcal{L}_{\mathsf{P}_{\geqslant r},\mathsf{T}}$ as $(\mathcal{L}_{\mathsf{P}_{\geqslant r}})_\mathsf{T}$, which will then contain predicates $\mathsf{P}_{\geqslant r}$ as well as a predicate T. In the syntax itself there is no mention of coding, a sentence of $\mathcal{L}_{\mathsf{P}_{\geqslant r}}$ might just be $\mathsf{P}_{\geqslant 1/2}\overline{3}$. The interpretation of this sentence as saying something about another sentence comes in via the coding of sentences. In the context of this language, then we assume that we have a coding which codes all sentences of $\mathcal{L}_{\mathsf{P}_{\geqslant r},\mathsf{T}}$. So there will be a sentence:

$$\mathsf{T}^\ulcorner\mathsf{P}_{=1}^{\ulcorner}\mathsf{T}^\ulcorner\mathsf{P}_{=1}^{\ulcorner}0 = 0^{\urcorner\urcorner\urcorner\urcorner}.$$

We can also use the coding to take care of these issues if we have one of these notions represented as a predicate and the other as an operator by defining the joint language to be the predicate language closed under the recursive definition of the operator language. E.g. then, we define the joint language $\mathcal{L}_{\mathsf{P}_{\geqslant r},\mathbb{T}}$ as $(\mathcal{L}_{\mathsf{P}_{\geqslant r}})_{\mathbb{T}}$. Then the coding that we use when considering this again codes all sentences, now of this language which includes an operator and some predicates.

Translations

Sometimes we are interested in translations between different languages. In such a translation we need to translate not only the outer-occurrences of the predicates but also occurrences "inside the codes". This will not work by an induction on formulas because there is no "deepest occurrence" of the predicate which the induction can start with. Instead the codings are shown to exist by using Kleene's recursion theorem.

Theorem 1.6.18 (Halbach 2014, Lemma 5.2)**.** *Consider recursive first-order languages $\mathcal{L}$ and $\mathcal{L}'$ where $\mathcal{L}$ contains the predicates Q_i, and $\mathcal{L}'$ contains $\mathcal{L}$ except that it may not contain the predicates Q_i and $\mathcal{L}'$ also extends $\mathsf{PA}_{\mathcal{L}'}$. Suppose $\psi_i(x)$ is a formula in $\mathcal{L}'$ which we will translate Q_i by.*

There is a translation function $\rho : \mathsf{Form}_{\mathcal{L}} \to \mathsf{Form}_{\mathcal{L}'}$ with:

$$
\rho(\varphi) = \begin{cases}
\psi_i(\underset{.}{\rho}(t)) & \varphi = Q_i t \\
\varphi & \varphi \text{ atomic and not of the form } Q_i t \\
\neg\rho(\psi) & \varphi = \neg\psi \\
\rho(\psi) \wedge \rho(\chi) & \varphi = \psi \wedge \chi \\
\forall x \rho(\psi) & \varphi = \forall x \psi
\end{cases}
$$

where $\underset{.}{\rho}$ is some object level formula representing ρ, so $\mathsf{PA}_{\mathcal{L}'} \vdash \underset{.}{\rho}(\ulcorner\varphi\urcorner) = \ulcorner\rho(\varphi)\urcorner$.

Using this one could show the facts alluded to about the relationships between the above languages.

1.7 Conditional probabilities

In this thesis we will not consider conditional probabilities and also do no work on determining an appropriate conditional for this language. Whenever we use $\varphi \to \psi$ in this thesis we mean the material implication that can also be defined as $\neg\varphi \vee \psi$. In this section, though, we will just mention one or two facts about conditional probabilities in this framework.

1.7.1 Deference is inconsistent

Just as the principle of introspection is inconsistent with probabilism, so too is a deference principle. Lewis's principle principle mentioned in the introduction is an example of a deference principle. The general structure of a deference principle says that some notion of probability p^{A}, for example an agent's subjective credences, should *defer* to another notion, p^{B}, for example the objective chances. So if p^{A} is "aware of" what probability value is assigned to φ by p^{B},

then p^A should agree with that assignment. We will here take the principle stating that A defers to B to imply:[28]

$$p^A(\varphi \mid P^B_{\geqslant 1/2}\ulcorner\varphi\urcorner) \geqslant 1/2$$
$$p^A(\varphi \mid \neg P^B_{\geqslant 1/2}\ulcorner\varphi\urcorner) \not\geqslant 1/2$$

Such a principle was first introduced by Miller (1966) and variants of it have been considered in many different situations with different interpretations. Van Fraassen's reflection principle from Van Fraassen (1984) takes P^A to be an agent's current probability and P^B her future ones. In Lewis's principal principle from Lewis (1980) Lewis interprets P^A as subjective probability and P^B as objective chance. In other places such a principle can express the fact that agent A takes agent B to be an expert. Finally if A = B then this formalises self-trust, a weakening of introspection.

Theorem 1.7.1. *Let $p^A : \mathcal{L}_{PB} \times \mathcal{L}_{PB} \to \mathbb{R}$ be a conditional probability function over arithmetic where $\mathcal{L}_{PB}$ is some first order language extending L_{PA} and contains the predicate symbol $P^B_{\geqslant 1/2}$. In particular, p^A should satisfy:*

- $p^A(\varphi \mid \varphi) = 1$ *if $p^A(\varphi) \neq 0$,*

- $p^A(\varphi \mid \neg\varphi) = 0$ *if $p^A(\neg\varphi) \neq 0$,*

- *If $p^A(\varphi) = 0$ then $p^A(\neg\varphi) \neq 0$,*

- *If $PA \vdash \varphi \leftrightarrow \psi$ and $p^A(\varphi) > 0$ and $p^A(\psi) > 0$, then $p^A(\chi \mid \varphi) = p^A(\chi \mid \psi)$.*

Then the following is unsatisfiable:

$$p^A(\varphi \mid P^B_{\geqslant 1/2}\ulcorner\varphi\urcorner) \geqslant 1/2$$
$$p^A(\varphi \mid \neg P^B_{\geqslant 1/2}\ulcorner\varphi\urcorner) \not\geqslant 1/2$$

Proof. Suppose $p^A(\pi) \neq 0$.

$$\begin{aligned}
1 &= p^A(\pi \mid \pi) \\
&= p^A(\pi \mid \neg P^B_{\geqslant 1/2}\ulcorner\pi\urcorner) \\
&\not\geqslant 1/2
\end{aligned}$$

This however does not deal with the case where $p^A(\pi) = 0$ since then, at least for some notions of conditional probability, $p^A(\pi \mid \pi)$ is undefined. However in that case $p^A(\neg\pi) = 1$, or at least $\neq 0$, so $p^A(\pi \mid \neg\pi)$ is well defined. Then:

$$\begin{aligned}
0 &= p^A(\pi \mid \neg\pi) \\
&= p^A(\pi \mid P^B_{\geqslant 1/2}\ulcorner\pi\urcorner) \\
&\geqslant 1/2 \qquad\qquad\qquad\qquad\qquad\qquad\qquad \square
\end{aligned}$$

---

[28]In fact that isn't implied from the previous principle in the introduction, which was $p^A(\varphi \mid P^B_{=r}\ulcorner\varphi\urcorner) = r$, but it does if instead we use a language like $\mathcal{L}^{\mathbb{R}}_{P_{\geqslant}}$ or $\mathcal{L}_{P_{\geqslant}}$ where one can quantify over the real numbers, or codes thereof.

So the principal principle, along with other analogous deference principles are inconsistent in this framework. It is known that so-called *undermining chances* can cause problems for the principal principle (see Lewis, 1994). An undermining chance is some function ch_0 such that it thinks another chance function is possible. π could be seen to cause every chance function to be undermining, since a non-undermining chance function would have to satisfy introspection, which we saw to be impossible in Section 1.4.

It would be interesting to study this result in more detail, but we do not do so in this thesis.

There are a number of further interesting questions that we would like to consider but which are not considered in this thesis. For example, which principles involving conditional probabilities are consistent? For example is

$$p(\varphi \mid \mathsf{T}\ulcorner\varphi\urcorner) = 1 \text{ and } p(\mathsf{T}\ulcorner\varphi\urcorner \mid \varphi) = 1$$

consistent? I would guess not, but this should be studied.

These are questions that are left to future research. One reason for this is that once higher order probabilities are allowed in the framework then the traditional understanding of conditional probabilities, perhaps by the ratio formula, seem to not appropriately formalise conditionalisation. So further analysis should be done to understand how to formulate conditionalisation at all and what we think it should satisfy before adding in extra principles involving truth or probability.

1.7.2 Why we won't consider them

In the language $\mathcal{L}_\mathsf{P}$, one can state the ratio formula and could therefore hope to just add a conditional probability function symbol defined by:

$$\mathsf{P}(\ulcorner\varphi\urcorner \mid \ulcorner\psi\urcorner) := \frac{\mathsf{P}(\ulcorner\varphi \wedge \psi\urcorner)}{\mathsf{P}(\ulcorner\psi\urcorner)} \qquad\qquad \text{if } \mathsf{P}(\ulcorner\psi\urcorner) > 0$$

One may therefore think that it would be a concept that would be easy to add to our framework. But unfortunately things aren't that simple. This is because the ratio definition does not work in the intended way when φ itself says something about probability.

Conditional probabilities are often used to capture *updating*. Under that understanding, a conditional subjective probability $p(\varphi \mid \psi)$ describes the degree of belief that the agent would have in φ if she were to learn ψ. We will show, that this is not given by the ratio formula.

We consider how intuitive this ratio formula is when we also have *higher order probabilities*. We will suggest that it in fact is not the appropriate formalisation of updating in such a language. This is *not* due to any self-referential probabilities like π as we can also observe this unintuitive feature in the operator language.

Example 1.7.2. Suppose a fair coin is about to be tossed. Let $\mathcal{L}$ be a propositional language with just one propositional variable *Heads* which is true if the coin will land heads. Consider the language $\mathcal{L}_{\mathbb{P}_{\geqslant r}}$ where the probability notion is supposed to refer to a particular agent's degrees of belief. And consider that agent's conditional degrees of belief for sentences of this language. So we consider $p : \mathcal{L}_{\mathbb{P}_{\geqslant r}} \times \mathcal{L}_{\mathbb{P}_{\geqslant r}} \to \mathbb{R}$ where p is supposed to be the same agent's degrees

of belief as the agent who is formalised in the language: she is reasoning about herself.

Suppose a fair coin has been tossed and that the agent has learnt that the coin landed heads. We might ask what her updated degree of belief that she is certain that *Heads* is. If the agent's conditional degree of belief captures what her belief is after learning, then this question can also be expressed by:

$$p(\mathbb{P}_{=1}(\textit{Heads}) \mid \textit{Heads}) =?$$

Intuitively, if the agent learns *Heads* she should be certain of *Heads* so should put $p(\textit{Heads}) = 1$. It would therefore be intuitive if we have

$$p(\mathbb{P}_{=1}(\textit{Heads}) \mid \textit{Heads}) = 1 \tag{1.6}$$

However, if we have a usual situation we would have that before the learning the agent has $p(\textit{Heads}) = 1/2$, and moreover, that she is certain of this fact. Therefore, remembering that p represents the agent's pre-learning degrees of belief we would have

$$p(\mathbb{P}_{=1/2}(\textit{Heads})) = 1$$

and therefore

$$p(\mathbb{P}_{=1}(\textit{Heads})) = 0$$

So the ratio formula would evaluate to:

$$\frac{p(\mathbb{P}_{=1}(\textit{Heads}) \wedge \textit{Heads})}{p(\textit{Heads})} = \frac{0}{p(\textit{Heads})} = 0 \tag{1.7}$$

But if the ratio formula appropriately described updating we would need that Eqs. (1.6) and (1.7) would evaluate to the same value.

This example will be made more formal and precise in Section 2.7.

Part I

Developing a Semantics

Chapter 2

Preliminaries and Challenges

The first part of this thesis focuses on developing a semantics for these languages with self-referential probabilities. A semantics will start off with some structures, or models, and say which sentences are true or false.

We will start with models of the language without the probability notion, and possibly some structure encoding certain facts about probability, and develop constraints on good extensions of the probability predicate over these. We will in fact use quite complicated background structures which will also embed certain information about how the probability notion modelled should work.

2.1 Probabilistic modal structures

As already discussed in Section 1.2.2, this thesis will have a particular focus on *subjective probability*, or the degrees of beliefs of agents, as the interpretation of probability. Since we may have different agents who have different degrees of belief we want to put these facts about the degrees of belief into the model itself. We do this by using possible world structures, in the form of *probabilistic modal structures*.

Many of our semantics will be based on such structures. Though not all of them, for example in Section 5.2 we will present a semantics that does not use such probabilistic modal structures. That semantics provides a different kind of notion of probability: one that measures semantic behaviour of sentences.

This use of probabilistic modal structures will allow many of the technical and conceptual advantages that probabilistic modal structures give us, as has been witnessed in the rise of modal logic (the analogous comment was made for the case of all-or-nothing modalities in Halbach et al. (2003)). We will answer the question of how such structures can be used to determine a semantics in these expressively rich languages.

2.1.1 What they are

Probabilistic modal structures are defined as follows.

Definition 2.1.1 (Probabilistic Modal Structure)**.** Fix some group of agents Agents. A *probabilistic modal structure,* $\mathfrak{M}$ for some language $\mathcal{L}$ is given by a frame and a valuation:

A *frame* is some $(W, \{m_w^A | w \in W, A \in \mathsf{Agents}\})$ where W is some non-empty set and m_w^A is some finitely additive probability measure over the powerset of W,[1] We call the m_w^A *accessibility measures.*

An *valuation,* $\mathbf{M}$ assigns to each world w, a classical model for the language $\mathcal{L}$. When we consider a language extending L_{PA} we will implicitly always assume that each $\mathbf{M}(w)$ is an $\mathbb{N}$-model.

We will use the un-boldface version of $\mathbf{M}$, M, for a single $\mathcal{L}$-model. More generally, we will use a boldface version of a symbol to denote something that assigns to each world one of the non-boldface items. For example, the metavariable for a function from sentences to reals that is being used as the extension of P is p, so we use $\mathbf{p}$ for a probabilistic evaluation function, which assigns to each world some such p. Occasionally we use $\boldsymbol{\mathcal{M}}$ for a function assigning to each world a model of the extended language, $\mathcal{M}$, and $\mathbf{T}$ to assign to each world an interpretation of T, T.

To see how these probabilistic modal structures work and how they are supposed to model particular set-ups consider the following example.

Example 2.1.2. Suppose we have an urn filled with 90 balls, 30 of which are yellow, 30 blue and 30 red. Suppose that a random ball is drawn from the urn and the agent is told whether it is yellow or not. We will give a probabilistic modal structure that represents the agent's degrees of belief once the ball has been drawn and she has been told whether it is yellow or not. To formalise this example we use a language, $\mathcal{L}$, with propositional variables *Yellow*, *Blue* and *Red*, which will stand for the propositions that a yellow, blue or red ball is drawn, respectively. We consider three worlds that will be used to represent the colour of the ball drawn, so we take $W = \{w_{Yellow}, w_{Blue}, w_{Red}\}$ where w_{Yellow} is actual if a yellow ball was drawn, w_{Blue} for the blue ball and w_{Red} for the red. The valuation function $\mathbf{M}$ describes what these worlds are like, for example the model $\mathbf{M}(w_{Yellow})$ assigns the truth-value true to *Yellow* and false to *Blue* and *Red*. The other component we need to finish our description of the probabilistic modal structure are the functions m_w representing how much our agent thinks the other worlds are possible if she is actually in the world w. If a yellow ball is actually drawn, i.e. the agent is in the world w_{Yellow}, then she is told that the ball is yellow, so she is certain that she is in w_{Yellow}. We therefore have that $m_{w_{Yellow}}(\{w_{Yellow}\}) = 1$, $m_{w_{Yellow}}(\{w_{Blue}\}) = 0$ and $m_{w_{Yellow}}(\{w_{Red}\}) = 0$. Since there are only finitely many worlds this is enough information to determine the full $m_{w_{Yellow}}$.[2] If a blue ball is drawn, i.e. the agent is in w_{Blue}, then she is told that the ball is not yellow so the only worlds she considers as still possible are w_{Blue} and w_{Red}. The agent thinks it is as likely that a blue ball is drawn as a red ball, so we will have that $m_{w_{Blue}}(\{w_{Yellow}\}) = 0$, $m_{w_{Blue}}(\{w_{Blue}\}) = 1/2$ and $m_{w_{Blue}}(\{w_{Red}\}) = 1/2$, which is again enough to determine the full $m_{w_{Blue}}$. The

---

[1]Assuming that this is defined on the whole powerset does not in fact lead to any additional restriction when we deal with merely-finitely additive probability measures, since a finitely additive probability measure on some Boolean algebra can always be extended to one defined on the whole powerset, see Proposition 1.2.2.

[2]Which will be given by: $m_{w_{Yellow}}(A) = \sum_{w \in A} m_{w_{Yellow}}(w)$.

case when a red ball is drawn is the same from the agent's perspective as if a blue ball is drawn so $m_{w_{Blue}} = m_{w_{Red}}$.

We can represent this probabilistic modal structure by the figure in Fig. 2.1.
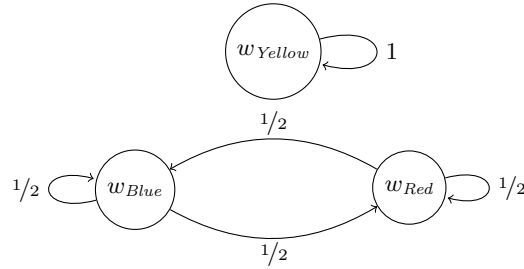
Figure 2.1: Representation of a finite probabilistic modal structure. Example 2.1.2.

In this example the space is finite so we can represent the measures by degree of accessibility relations, which we have done by the labelled arrows in the diagram. We have omitted the the arrows that would be labelled by 0.

If the frame is finite and the agent(s) are introspective[3] then we can further simplify the diagram representing this probabilistic modal structure. For example, as follows:
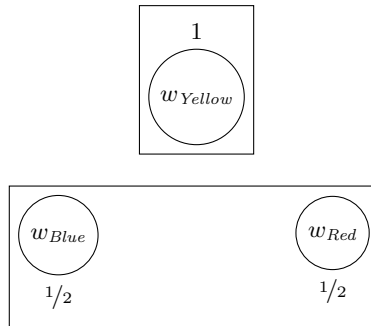
Figure 2.2: Since Example 2.1.2 is finite and the agent is introspective we can also represent it this way.

To see how this diagram works, one has that m_w is determined by the weights of the worlds in the box in which m_w is a part. For any v not in the box containing w we put $m_w\{v\} = 0$.

In the previous example, each of the worlds had a different "state of affairs" assigned to it, given by the **M**. But this is not always true. Consider for example the following probabilistic modal structure:

Example 2.1.3. A fair coin is tossed and Suzanne is not told the result of the toss. Suzanne isn't sure whether Ed has been told the result of the toss or not.

---

[3]By which we mean that the frame is strongly introspective. See Section 2.3.1 for a formal definition.

She thinks it is equally likely that he has been told that he has not. We would represent this as in Fig. 2.3
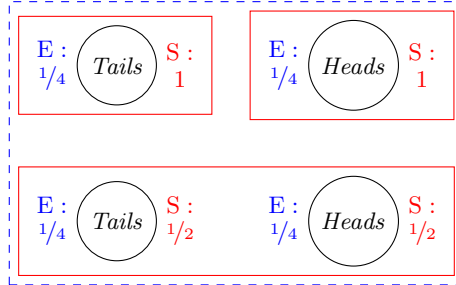
Figure 2.3: The red (solid) information on this diagram captures Suzanne's accessibility measure, the blue (dashed) information captures Ed's.

There is one particularly interesting probabilistic modal structure which we will consider throughout this thesis:

Definition 2.1.4. $\mathfrak{M}$ is *omniscient* if W is a singleton.
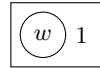
Figure 2.4: $\mathfrak{M}_{\mathsf{omn}}$.

Let $\mathfrak{M}_{\mathsf{omn}}$ denote some omniscient probabilistic modal structure.

In the finite case these are easy generalisations of Kripke models (as studied in modal logic), where the accessibility relation is replaced by degree-of-accessibility relations. In the infinite case the accessibility relations from Kripke models are replaced by general accessibility *measures* which have information about how w is connected to a set of worlds as well as how it is connected to each individual world v. I present these in a way that is analogous to the Kripke models because my use of them to determine semantics for a language with self-referential probabilities is closely connected to work in Halbach et al. (2003); Halbach and Welch (2009); Stern (2015b, 2014a,b) where they use Kripke models to consider modality conceived of as a predicate.

Except for some details, these are the same as what are called Aumann structures following Aumann (1999), and are also the same as type spaces, at least in some of the presentations of such spaces, e.g. Heifetz and Mongin (2001). In the type-spaces literature the accessibility measures are almost always assumed to be countably additive and it is also often assumed that the structures are introspective, i.e. that

$$m_w\{v \mid m_v = m_w\} = 1,$$

though those would generally be called a *Harysani type space*. The type-spaces literature typically just studies the *universal type-space*, which is a type-space into which any other can be embedded. However, the existence of such a universal type-space requires the assumption of σ-additivity and don't exist if we only assume finite additivity (Meier, 2006).

We do not assume that the structures are countably additive for a few reasons:

Firstly, there are many arguments that subjective probabilities should be finitely additive, and finite additivity is supported by many authors, but the extension to being countably additive has weaker justifications and has been rejected by a number of authors, most notably de Finetti. By assuming countably additivity one restricts the possible probability measures, for example there are no countably additive uniform measures over the natural numbers and no extension of the Lebesgue measure that measures *all* subsets of the unit interval.

Furthermore, we will show in Section 2.4.2 that the assumption of countable additivity rules out the possibility of using the structure in the intuitive way to assign probabilities to sentences, because there are then no *Prob-PW-models*, whereas that hasn't been ruled out if we only assume finite additivity. Since finite additivity may be appropriate for representing subjective probabilities it would be interesting to determine if any (merely-finitely-additive) structures do support Prob-PW-models.

We in fact assumed that the accessibility measure was defined on the whole powerset, which can be done with finitely additive measures but not generally with countably additive measures unless the underlying space is finite. Much of what is said in this thesis can be said without the assumption that the accessibility measure be defined on the whole powerset, it is generally used just as a technical assumption that makes the presentation easier. However, in the fixed point semantics, e.g. Chapter 3, the assumption is required if one wishes to study arbitrary fixed points, as these require arbitrary subsets to have measures attached to them. For example our axiomatisation characterises the collection of fixed points, so we have to allow any of these fixed points to be well-defined. If instead one is only interested in the minimal fixed point, then one could provide an account of which sets are required to be measurable, but since we are interested in more than that we do not restrict the measurable sets.

These probabilistic modal structures allow us to define an easy and natural semantics for an operator language.

2.2 Semantics for an operator language using probabilistic modal structures

These probabilistic modal structures can very simply be used to give a semantics for an *operator* language for probability. We presented the language in Definition 1.6.12 which contains probability operators $\mathbb{P}_{\geqslant r}$ for each rational r.

We can now define which of such sentences are true or false at different worlds in a given probabilistic modal structure.

Definition 2.2.1 (Semantics for $\mathcal{L}_{\mathbb{P}}$)**.** Fix some probabilistic modal structure $\mathfrak{M}$. Define:

- For $\varphi \in \mathsf{Sent}_{\mathcal{L}}$, $w \models_{\mathfrak{M}} \varphi \iff \mathbf{M}(w) \models \varphi$

- $w \models_{\mathfrak{M}} \varphi \vee \psi \iff w \models_{\mathfrak{M}} \varphi$ or $w \models_{\mathfrak{M}} \psi$

- $w \models_{\mathfrak{M}} \neg \varphi \iff w \not\models_{\mathfrak{M}} \varphi$

- $w \models_{\mathfrak{M}} \mathbb{P}_{\geqslant r} \varphi \iff m_w \{v \mid v \models_{\mathfrak{M}} \varphi\} \geqslant r$

- $w \models_{\mathfrak{M}} \mathbb{P}_{>r}\varphi \iff m_w\{v \mid v \models_{\mathfrak{M}} \varphi\} > r$

As in Definition 1.6.13 we can also talk about the other variant operators like $\mathbb{P}_{\leqslant r}$. The way we defined these was good because it satisfies the following.

Proposition 2.2.2.

$$w \models_{\mathfrak{M}} \mathbb{P}_{\leqslant r}\varphi \iff m_w\{v \mid v \models_{\mathfrak{M}} \varphi\} \leqslant r$$
$$w \models_{\mathfrak{M}} \mathbb{P}_{=r}\varphi \iff m_w\{v \mid v \models_{\mathfrak{M}} \varphi\} = r$$
$$w \models_{\mathfrak{M}} \mathbb{P}_{<r}\varphi \iff m_w\{v \mid v \models_{\mathfrak{M}} \varphi\} < r$$

Proof. We only write the proof for the first property:

Remember we defined $\mathbb{P}_{\leqslant r}\varphi$ as $\mathbb{P}_{\geqslant 1-r}\neg\varphi$. So by using the definition of the semantics for $\mathbb{P}_{\geqslant r}\neg\varphi$ it suffices to show:

$$m_w\{v \mid v \models_{\mathfrak{M}} \varphi\} = 1 - m_w\{v \mid v \models_{\mathfrak{M}} \neg\varphi\}$$

because then

$$m_w\{v \mid v \models_{\mathfrak{M}} \varphi\} \leqslant r \iff m_w\{v \mid v \models_{\mathfrak{M}} \neg\varphi\} \geqslant 1 - r.$$

Observe that $\{\{v \mid v \models_{\mathfrak{M}} \varphi\}, \{v \mid v \models_{\mathfrak{M}} \neg\varphi\}\}$ is a partition of W. So since m_w is finitely additive, we have

$$m_w\{v \mid v \models_{\mathfrak{M}} \varphi\} + m_w\{v \mid v \models_{\mathfrak{M}} \neg\varphi\} = 1,$$

as required. $\qquad\square$

One can similarly give a semantics for the other considered operator languages. This semantics gives us some idea of how these probabilistic modal structures are supposed to work.

In Section 2.4 we will discuss how one might use these structures to give a semantics in the case where probability is formalised in the expressively rich setting, i.e. as a predicate or function symbol. But first we will consider what assumptions are being made by using these probabilistic modal structures and identify some particularly interesting (classes of) probabilistic modal structures.

2.3 Assumptions in probabilistic modal structures

We have assumed that the m_w^{A} are all probability measures. This means that agents who are representable in such structures must be probabilistically coherent. Also, they must be certain that all other agents are probabilistically coherent. And so on. So they must have common certainty of probabilistic coherence.

We can also add additional assumptions to narrow down the class of probabilistic modal structures that are studied.

2.3.1   Introspective structures

We have already discussed introspection as a principle

$$\mathbb{P}_{\geqslant r}\varphi \to \mathbb{P}_{=1}\mathbb{P}_{\geqslant r}\varphi$$
$$\neg\mathbb{P}_{\geqslant r}\varphi \to \mathbb{P}_{=1}\neg\mathbb{P}_{\geqslant r}\varphi$$

Which, in the predicate case, was inconsistent with probabilism. We can characterise the probabilistic modal structures which satisfy these introspection principles in the operator case.

Definition 2.3.1. A probabilistic modal frame $(W, \{m_w \,|\, w \in W\})$, or structure $\mathfrak{M}$, is *strongly introspective* if for all $w \in W$

$$m_w\{v \,|\, m_v = m_w\} = 1$$

A probabilistic modal frame $(W, \{m_w \,|\, w \in W\})$, or structure $\mathfrak{M}$, is *weakly introspective* if for all $w \in W$, $A \subseteq W$ and $r \leqslant m_w(A) < q$,

$$m_w\{v \,|\, r \leqslant m_v(A) < q\} = 1$$

If the frame is countably additive then strong and weak introspection are equivalent. However, if the frame is merely-finely-additive, then they may not be equivalent, though being strongly introspective implies being weakly introspective. Being strongly introspective is the standard definition, for example as is found in Heifetz and Mongin (2001), due to the fact that usually only countably additive frames are considered.

Proposition 2.3.2. *Let $\mathcal{L}$ be any propositional or first order language with at least one propositional variable, and $\mathcal{L}_{\mathbb{P}_{\geqslant r}}$ as defined in Definition 1.6.12.*
A probabilistic modal frame for $\mathcal{L}_{\mathbb{P}_{\geqslant r}}$, $(W, \{m_w \,|\, w \in W\})$ is weakly introspective iff for every $\mathfrak{M}$ based on $(W, \{m_w \,|\, w \in W\})$ and for each $w \in W$,

$$w \models_{\mathfrak{M}} \mathbb{P}_{\geqslant r}\varphi \to \mathbb{P}_{=1}\mathbb{P}_{\geqslant r}\varphi$$
$$w \models_{\mathfrak{M}} \neg\mathbb{P}_{\geqslant r}\varphi \to \mathbb{P}_{=1}\neg\mathbb{P}_{\geqslant r}\varphi$$

Proof. $\implies$: Suppose $\mathfrak{M}$ is based on a weakly introspective frame. Then

$$
\begin{aligned}
w \models \mathbb{P}_{\geqslant r}\varphi \iff{}& m_w\{v \,|\, v \models_{\mathfrak{M}} \varphi\} \geqslant r \\
\implies{}& m_w\{v' \,|\, m_{v'}\{v \,|\, v \models_{\mathfrak{M}} \varphi\} \geqslant r\} = 1 \\
\iff{}& m_w\{v' \,|\, v' \models_{\mathfrak{M}} \mathbb{P}_{\geqslant r}\varphi\} = 1 \\
\iff{}& w \models \mathbb{P}_{=1}\mathbb{P}_{\geqslant r}\varphi \\
w \models \neg\mathbb{P}_{\geqslant r}\varphi \iff{}& m_w\{v \,|\, v \models_{\mathfrak{M}} \varphi\} < r \\
\implies{}& m_w\{v' \,|\, m_{v'}\{v \,|\, v \models_{\mathfrak{M}} \varphi\} < r\} = 1 \\
\iff{}& m_w\{v' \,|\, v' \models_{\mathfrak{M}} \neg\mathbb{P}_{\geqslant r}\varphi\} = 1 \\
\iff{}& w \models \neg\mathbb{P}_{=1}\mathbb{P}_{\geqslant r}\varphi
\end{aligned}
$$

For $\impliedby$: suppose the RHS. Fix w and some $A \subseteq W$. Suppose $\mathcal{L}$ contains the propositional variable O. Consider a valuation $\mathbf{M}$ such that

$$\mathbf{M}(v) \models O \iff v \in A.$$

Suppose $r \leqslant m_w(A) < q$. Then $w \models_{\mathfrak{M}} \mathbb{P}_{\geqslant r} O$ and $w \models_{\mathfrak{M}} \neg \mathbb{P}_{\geqslant q} O$. So $w \models_{\mathfrak{M}} \mathbb{P}_{=1} \mathbb{P}_{\geqslant r} O$ and $w \models_{\mathfrak{M}} \mathbb{P}_{=1} \neg \mathbb{P}_{\geqslant q} O$, i.e. $m_w\{v \mid r \leqslant m_v(A)\} = 1$ and $m_w\{v \mid m_v(A) < q\} = 1$. Since m_w is finitely additive, we also have:

$$m_w\{v \mid r \leqslant m_v(A) < q\} = 1 \qquad \square$$

Let us now start to look at the version where we don't work with a probability operator but instead a probability *predicate*.

2.4   Semantics in the predicate case

2.4.1   What is a Prob-PW-model

In Section 2.2 we have seen a semantics in the operator language. The natural semantics to try to define for the predicate language would just work by applying the same operator clauses but now to the predicate language.

We will work with $\mathcal{L}_{\mathsf{P}_{\geqslant r}}$, that extends any $\mathcal{L}$ with the predicates $\mathsf{P}_{\geqslant r}$ for each rational r.

Setup 2 (Section 2.4)**.** *Let $\mathcal{L}$ be some recursive first order language extending L_{PA}. We will work with $\mathcal{L}_{\mathsf{P}_{\geqslant r}}$ as described in Definition 1.6.7.*

We can attempt to define the semantics for $\mathcal{L}_{\mathsf{P}_{\geqslant r}}$ as we did for the operator case in Definition 2.2.1 by the clauses:

- For $\varphi \in \mathsf{Sent}_{\mathcal{L}}$, $w \models_{\mathfrak{M}} \varphi \iff \mathbf{M}(w) \models \varphi$

- $w \models_{\mathfrak{M}} \varphi \vee \psi \iff w \models_{\mathfrak{M}} \varphi$ or $w \models_{\mathfrak{M}} \psi$

- $w \models_{\mathfrak{M}} \neg \varphi \iff w \not\models_{\mathfrak{M}} \varphi$

- $w \models_{\mathfrak{M}} \mathsf{P}_{\geqslant r} \ulcorner \varphi \urcorner \iff m_w\{v \mid v \models_{\mathfrak{M}} \varphi\} \geqslant r$,

- $w \models_{\mathfrak{M}} \mathsf{P}_{> r} \ulcorner \varphi \urcorner \iff m_w\{v \mid v \models_{\mathfrak{M}} \varphi\} > r$.

In the operator case we worked with propositional logic but in this predicate language we are dealing with predicate logic so we want to add an extra clause dealing with the quantifiers as well as Leibnitz's law.[4]

- $w \models_{\mathfrak{M}} \exists x \varphi(x) \iff$ for some $n \in \mathbb{N}$, $w \models_{\mathfrak{M}} \varphi(\overline{n})$,

- $w \models_{\mathfrak{M}} t = s \implies (w \models_{\mathfrak{M}} \varphi(t) \iff w \models_{\mathfrak{M}} \varphi(s))$.

The problem with trying to use this as a definition is that we don't have a recursive specification of the formulas: $\mathsf{P}_{\geqslant r} \ulcorner \varphi \urcorner$ is just some $\mathsf{P}_{\geqslant r} \overline{n}$ and is an atomic formula, however complex φ is. So this definition is not a recursive definition and is therefore we are not guaranteed to have any $\models_{\mathfrak{M}}$ that satisfies the attempted definition. And in fact it does often turn out to be unsatisfiable, as we will now show.

We will first show that if these clauses hold for $\models_{\mathfrak{M}}$ then for each w, there is a first order model that satisfies the same sentences as are satisfied at w according $\models_{\mathfrak{M}}$. Furthermore, these first order models take a particular form: the $\mathcal{L}$-component of the first order model is $\mathbf{M}(w)$, and the interpretation of each of the probability predicates can be given by a prob-eval function.

---

[4]Instead of the general Leibnitz's law, we could have just included the instances for φ of the form $\mathsf{P}_{\geqslant r} x$ and $\mathsf{P}_{> r} x$.

Definition 2.4.1. A *prob-eval function,* $\mathbf{p}$, assigns to each world a function from $\mathsf{Sent}_{\mathsf{P}_{\geqslant r}}$ to $\mathbb{R}$.

And we will use this to give a first-order model of $\mathcal{L}_{\mathsf{P}_{\geqslant r}}$ at each world by:

Definition 2.4.2. $(\mathbf{M}, \mathbf{p})(w) = (\mathbf{M}(w), \mathbf{p}(w))$ is the expansion of the $\mathcal{L}$-model, $\mathbf{M}(w)$, to a $\mathcal{L}_{\mathsf{P}_{\geqslant r}}$-model by

$$(\mathbf{M}, \mathbf{p})(w) \models \mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner \iff \mathbf{p}(w)(\varphi) \geqslant r$$
$$(\mathbf{M}, \mathbf{p})(w) \models \mathsf{P}_{> r}\ulcorner\varphi\urcorner \iff \mathbf{p}(w)(\varphi) > r$$

If there is no $\varphi \in \mathsf{Sent}_{\mathsf{P}_{\geqslant r}}$ with $n = \#\varphi$, then

$$(\mathbf{M}, \mathbf{p})(w) \models \mathsf{P}_{\geqslant r}\overline{n} \iff r \leqslant 0$$
$$(\mathbf{M}, \mathbf{p})(w) \models \mathsf{P}_{> r}\overline{n} \iff r < 0$$

In this, we take probability 0 to be the default value, so for all n, $(\mathbf{M}, \mathbf{p})(w) \models \mathsf{P}_{\geqslant 0}\overline{n}$. We do this to allow a smooth transition to the semantics we'll provide in Chapter 3, though this choice doesn't make any essential difference.

Specifically, then, we will show that if $\models_{\mathfrak{M}}$ satisfies these clauses then there is some prob-eval function, $\mathbf{p}$, where $w \models_{\mathfrak{M}} \varphi \iff (\mathbf{M}, \mathbf{p})(w) \models \varphi$. In fact this first result hasn't yet said much about the interpretation of the probability predicate, and this can be seen by the fact that we don't in fact need $\models_{\mathfrak{M}}$ to satisfy the probability clauses to get this result, instead they only need to satisfy some minimal clauses, for example we require that $w \models_{\mathfrak{M}} \mathsf{P}_{\geqslant 1/2}\ulcorner\varphi\urcorner \implies w \models_{\mathfrak{M}} \mathsf{P}_{\geqslant 1/4}\ulcorner\varphi\urcorner$.

Proposition 2.4.3. *Let $\mathfrak{M}$ be such that each $\mathbf{M}(w)$ is an $\mathbb{N}$-model.*
$\models_{\mathfrak{M}} \subseteq W \times \mathsf{Sent}_{\mathsf{P}_{\geqslant r}}$ satisfies:

- *For $\varphi \in \mathsf{Sent}_{\mathcal{L}}$, $w \models_{\mathfrak{M}} \varphi \iff \mathbf{M}(w) \models \varphi$*

- *$w \models_{\mathfrak{M}} \varphi \vee \psi \iff w \models_{\mathfrak{M}} \varphi$ or $w \models_{\mathfrak{M}} \psi$*

- *$w \models_{\mathfrak{M}} \neg\varphi \iff w \not\models_{\mathfrak{M}} \varphi$*

- *$w \models_{\mathfrak{M}} \exists x \varphi(x) \iff$ for some $n \in \mathbb{N}$, $w \models_{\mathfrak{M}} \varphi(\overline{n})$*

- *$w \models_{\mathfrak{M}} t = s \implies (w \models_{\mathfrak{M}} \varphi(t) \iff w \models_{\mathfrak{M}} \varphi(s))$.*

and is such that

- *$w \models_{\mathfrak{M}} \mathsf{P}_{\geqslant r}\overline{n} \iff$ for all $q < r$, $w \models_{\mathfrak{M}} \mathsf{P}_{\geqslant q}\overline{n}$,*

- *$w \models_{\mathfrak{M}} \mathsf{P}_{> r}\overline{n} \iff$ there is some $q > r$, such that $w \models_{\mathfrak{M}} \mathsf{P}_{\geqslant q}\overline{n}$,*

- *If there is no $\varphi \in \mathsf{Sent}_{\mathsf{P}_{\geqslant r}}$ with $n = \#\varphi$ then $w \models_{\mathfrak{M}} \mathsf{P}_{\geqslant 0}\overline{n} \wedge \neg\mathsf{P}_{> 0}\overline{n}$.*

if and only if there is some prob-eval function $\mathbf{p}$, such that

$$w \models_{\mathfrak{M}} \varphi \iff (\mathbf{M}, \mathbf{p})(w) \models \varphi.$$

Proof. $\Longleftarrow$: We need to show that $(\mathbf{M}, \mathbf{p})(w)$ satisfies the relevant clauses. The only interesting ones are those involving probability,[5] which use the density of $\mathbb{Q}$ in $\mathbb{R}$.

$\Longrightarrow$: Define $\mathbf{p}(w)(\varphi) = \sup\{r \mid w \models_{\mathfrak{M}} \mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner\}$. One can show

$$w \models_{\mathfrak{M}} \varphi \iff (\mathbf{M}, \mathbf{p})(w) \models \varphi$$

by induction on the complexity of φ.[6] $\qquad\square$

Choosing such a $\mathbf{p}$ allows $\models_{\mathfrak{M}}$ to satisfy the attemped definition clauses for the connectives, but it is not yet guaranteed that it appropriately interprets the probability predicates. To achieve that, we need $\mathbf{p}$ to also satisfy:

$$(\mathbf{M}, \mathbf{p})(w) \models \mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner \iff m_w\{v \mid (\mathbf{M}, \mathbf{p})(v) \models \varphi\} \geqslant r,$$
$$(\mathbf{M}, \mathbf{p})(w) \models \mathsf{P}_{> r}\ulcorner\varphi\urcorner \iff m_w\{v \mid (\mathbf{M}, \mathbf{p})(v) \models \varphi\} > r.$$

We ensure that $\mathbf{p}$ does this by turning the desired clause into a definition of an operator taking $\mathbf{p}$ to another prob-eval function.

Definition 2.4.4. Let $\mathbf{p}$ be any prob-eval function. Define the prob-eval function $\Theta_{\mathfrak{M}}(\mathbf{p})$ by:

$$\Theta_{\mathfrak{M}}(\mathbf{p})(w)(\varphi) := m_w\{v \mid (\mathbf{M}, \mathbf{p})(v) \models \varphi\}.$$

We need to pick some $\mathbf{p}$ with $\Theta_{\mathfrak{M}}(\mathbf{p}) = \mathbf{p}$, as it will then satisfy the relevant clauses.

Proposition 2.4.5. $\mathbf{p}$ *is such that*

$$(\mathbf{M}, \mathbf{p})(w) \models \mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner \iff m_w\{v \mid (\mathbf{M}, \mathbf{p})(v) \models \varphi\} \geqslant r,$$
$$(\mathbf{M}, \mathbf{p})(w) \models \mathsf{P}_{> r}\ulcorner\varphi\urcorner \iff m_w\{v \mid (\mathbf{M}, \mathbf{p})(v) \models \varphi\} > r.$$

if and only if $\Theta_{\mathfrak{M}}(\mathbf{p}) = \mathbf{p}$.

So choosing such a fixed point $\mathbf{p}$ is essentially picking some interpretation of $\models_{\mathfrak{M}}$ that satisfies the desired defining clauses.

Corollary 2.4.6. *Let* $\mathfrak{M}$ *be such that each* $\mathbf{M}(w)$ *is an* $\mathbb{N}$*-model.*
$\models_{\mathfrak{M}} \subseteq W \times \mathsf{Sent}_{\mathsf{P}_{\geqslant r}}$ *satisfies:*

- *For* $\varphi \in \mathcal{L}$, $w \models_{\mathfrak{M}} \varphi \iff \mathbf{M}(w) \models \varphi$

- $w \models_{\mathfrak{M}} \varphi \lor \psi \iff w \models_{\mathfrak{M}} \varphi$ *or* $w \models_{\mathfrak{M}} \psi$

- $w \models_{\mathfrak{M}} \neg\varphi \iff w \not\models_{\mathfrak{M}} \varphi$

- $w \models_{\mathfrak{M}} \exists x \varphi(x) \iff$ *for some* $n \in \mathbb{N}$, $w \models_{\mathfrak{M}} \varphi(\overline{n})$,

- $w \models_{\mathfrak{M}} t = s \implies (w \models_{\mathfrak{M}} \varphi(t) \iff w \models_{\mathfrak{M}} \varphi(s))$.

- $w \models_{\mathfrak{M}} \mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner \iff m_w\{v \mid v \models_{\mathfrak{M}} \varphi\} \geqslant r$,

---

[5]The $\Longleftarrow$ of the $\geqslant$ clause and the $\Longrightarrow$ of the $>$ clause.

[6]For the base cases $\varphi = \mathsf{P}_{\geqslant r}t$ and $\mathsf{P}_{> r}t$ we use the fact that there will be some n such that $\mathbf{M}(w) \models t = \overline{n}$, so $(\mathbf{M}, \mathbf{p})(w) \models t = \overline{n}$ and $w \models_{\mathfrak{M}} t = \overline{n}$.

- $w \models_{\mathfrak{M}} \mathsf{P}_{>r}\ulcorner\varphi\urcorner \iff m_w\{v \mid v \models_{\mathfrak{M}} \varphi\} > r$.

- *If there is no* $\varphi \in \mathsf{Sent}_{\mathsf{P}_{\geqslant r}}$ *with* $n = \#\varphi$ *then* $w \models_{\mathfrak{M}} \mathsf{P}_{\geqslant 0}\overline{n} \wedge \neg\mathsf{P}_{>0}\overline{n}$.

if and only if there is some prob-eval function, $\mathbf{p}$, *such that*

$$\Theta_{\mathfrak{M}}(\mathbf{p}) = \mathbf{p},$$

and:

$$w \models_{\mathfrak{M}} \varphi \iff (\mathbf{M}, \mathbf{p})(w) \models \varphi.$$

Proof. This follows from Propositions 2.4.3 and 2.4.5 using the fact that

- $w \models_{\mathfrak{M}} \mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner \iff m_w\{v \mid v \models_{\mathfrak{M}} \varphi\} \geqslant r$, and
- $w \models_{\mathfrak{M}} \mathsf{P}_{>r}\ulcorner\varphi\urcorner \iff m_w\{v \mid v \models_{\mathfrak{M}} \varphi\} > r$,

imply

- $w \models_{\mathfrak{M}} \mathsf{P}_{\geqslant r}\overline{n} \iff$ for all $q < r$, $w \models_{\mathfrak{M}} \mathsf{P}_{\geqslant q}\overline{n}$, and
- $w \models_{\mathfrak{M}} \mathsf{P}_{>r}\overline{n} \iff$ there is some $q > r$, such that $w \models_{\mathfrak{M}} \mathsf{P}_{\geqslant q}\overline{n}$,

due to the density of $\mathbb{Q}$ in $\mathbb{R}$. $\qquad\square$

Such fixed point evaluation functions therefore allow the intuitive semantics clauses to be applied consistently.

Note that $\Theta_{\mathfrak{M}}(\mathbf{p})$ is probabilistic.

Definition 2.4.7. A probabilistic modal structure $\mathfrak{M}$ *supports a Prob-PW-model* if there is some prob-eval function $\mathbf{p}$ on $\mathfrak{M}$ with $\mathbf{p} = \Theta_{\mathfrak{M}}(\mathbf{p})$.

A probabilistic modal frame $(W, \{m_w \mid w \in W\})$ *supports a Prob-PW-model* if there is some $\mathbf{M}$ on $(W, \{m_w \mid w \in W\})$ where $\mathfrak{M} = (W, \{m_w \mid w \in W\}, \mathbf{M})$ supports a Prob-PW-model.

We will show that not all probabilistic modal structures do support Prob-PW-models. Due to the result in Proposition 2.4.3 that says that it is exactly the Prob-PW-models that satisfy the attempted definition of $\models_{\mathfrak{M}}$, this claim is ultimately the same as that made at the beginning of this section: that there are cases where the desired clauses for the semantics are not satisfiable. It would be interesting to answer the further question of which probabilistic modal structures *do* support Prob-PW-models, but this thesis will not answer that. Instead we just identify certain classes of structures that *do not* support a Prob-PW-model. In particular we will show this for any introspective, finite, or countably additive structures.

2.4.2 Not all probabilistic modal structures support Prob-PW-models

$\mathfrak{M}_{\mathsf{omn}}$ does not support Prob-PW-models

Some probabilistic modal structures will not support Prob-PW-models. For example:

Theorem 2.4.8. $\mathfrak{M}_{\mathsf{omn}}$ *does not support a Prob-PW-model.*
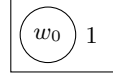
Figure 2.5: $\mathfrak{M}_{\mathsf{omn}}$.

Proof. Suppose $\mathbf{p}$ were a Prob-PW-model on $\mathfrak{M}_{\mathsf{omn}}$. Consider

$$\mathsf{PA} \vdash \pi \leftrightarrow \neg\mathsf{P}_{\geqslant 1/2}\ulcorner \pi \urcorner.$$

$$
\begin{aligned}
(\mathbf{M},\mathbf{p})(w_0) \models \pi &\implies m_{w_0}\{v \mid (\mathbf{M},\mathbf{p})(v) \models \pi\} = 1 \\
&\implies (\mathbf{M},\mathbf{p})(w_0) \models \mathsf{P}_{=1}\ulcorner \pi \urcorner \\
&\implies (\mathbf{M},\mathbf{p})(w_0) \models \mathsf{P}_{\geqslant 1/2}\ulcorner \pi \urcorner \\
&\implies (\mathbf{M},\mathbf{p})(w_0) \models \neg\pi \\
(\mathbf{M},\mathbf{p})(w_0) \models \neg\pi &\implies m_{w_0}\{v \mid (\mathbf{M},\mathbf{p})(v) \models \pi\} = 0 \\
&\implies (\mathbf{M},\mathbf{p})(w_0) \models \mathsf{P}_{=0}\ulcorner \pi \urcorner \\
&\implies (\mathbf{M},\mathbf{p})(w_0) \models \neg\mathsf{P}_{\geqslant 1/2}\ulcorner \pi \urcorner \\
&\implies (\mathbf{M},\mathbf{p})(w_0) \models \pi \qquad\qquad \square
\end{aligned}
$$

In this case we essentially have the liar paradox again. Due to the setup, and as was described in this proof, we would have that

$$(\mathbf{M},\mathbf{p})(w_0) \models \varphi \iff (\mathbf{M},\mathbf{p})(w_0) \models \mathsf{P}_{=1}\ulcorner \varphi \urcorner,$$

so $\mathsf{P}_{=1}$ acts like a truth predicate that satisfies the T-biconditionals, which we know not to be possible due to the liar paradox.

Introspective structures do not support Prob-PW-models

Introspective frames do not support a Prob-PW-model as a direct consequence of Theorem 1.4.1 which said that introspection and probabilism are inconsistent. This is just another way to view that result.

Theorem 2.4.9. *Suppose $(W, \{m_w \mid w \in W\})$ is weakly introspective. Then it does not support a Prob-PW-model.*

Proof. We prove this by means of a lemma:

Lemma 2.4.9.1. *Suppose $\mathfrak{M}$ is weakly introspective and $\mathbf{p} = \Theta_{\mathfrak{M}}(\mathbf{p})$. Then*

$$
(\mathbf{M},\mathbf{p})(w) \models
\begin{array}{l}
\mathsf{P}_{\geqslant r}\ulcorner \varphi \urcorner \rightarrow \mathsf{P}_{=1}\ulcorner \mathsf{P}_{\geqslant r}\ulcorner \varphi \urcorner \urcorner \\
\wedge \quad \neg\mathsf{P}_{\geqslant r}\ulcorner \varphi \urcorner \rightarrow \mathsf{P}_{=1}\ulcorner \neg\mathsf{P}_{\geqslant r}\ulcorner \varphi \urcorner \urcorner \\
\wedge \quad \mathsf{Prov}_{\mathsf{PA}}(\ulcorner \varphi \leftrightarrow \psi \urcorner) \rightarrow (\mathsf{P}_{\geqslant r}\ulcorner \varphi \urcorner \leftrightarrow \mathsf{P}_{\geqslant r}\ulcorner \psi \urcorner) \\
\wedge \quad \mathsf{P}_{=1}\ulcorner \varphi \urcorner \rightarrow \neg\mathsf{P}_{\geqslant 1/2}\ulcorner \neg\varphi \urcorner
\end{array}
$$

Proof. Showing that the introspection principles are satisfied is directly analogous to in Proposition 2.3.2. The other two are direct consequences of $\mathbf{p}(w)$ being probabilistic over PA. $\qquad\square$

Observe that in Lemma 2.4.9.1 we showed if $\mathfrak{M}$ is weakly introspective and $\mathbf{p} = \Theta_{\mathfrak{M}}(\mathbf{p})$, then $(\mathbf{M},\mathbf{p})(w)$ satisfies a theory which, in Theorem 1.4.1, we showed to be inconsistent. $\qquad\square$

Finite and countably additive frames do not support Prob-PW-models

There is a further large swathe of frames that do not support Prob-PW-models.

Theorem 2.4.10. *If $\mathfrak{M}$ is a probabilistic modal structure where each m_w is countably additive and each $\mathbf{M}(w)$ is an $\mathbb{N}$-model, then $\mathfrak{M}$ does not support a Prob-PW-model.*

Therefore, if $\mathfrak{M}$ is a finite probabilistic modal structure and each $\mathbf{M}(w)$ is an $\mathbb{N}$-model, then it does not support a Prob-PW-model.

This will be because if there were a Prob-PW-model, then the agent would have fully introspected certainty of being probabilistic and $\mathbb{N}$-additive. But that will turn out not to be possible because of a very influential result from McGee, (McGee, 1985), involving a sentence

(γ) I do not have fully introspected certainty of γ.

The idea of the challenge for countable additivity as a result of this theorem was given to me by Hannes Leitgeb.

We will use $\mathbb{N}\mathsf{AddPr}$ (for "$\mathbb{N}$-additive probability") to say that $\mathsf{P}_{=1}$ satisfies the properties required to come from some $\mathbb{N}$-additive probability function P (over Peano arithmetic, PA).[7] $\mathsf{IntroNAddPr}$ (for "introspected $\mathbb{N}$-additive probability") will say something like that P has fully introspected certainty that it is an $\mathbb{N}$-additive probability function.

Definition 2.4.11. Let $\mathbb{N}\mathsf{AddPr}$ denote the theory consisting of all instances of the following schema for any $\varphi, \psi \in \mathsf{Sent}_{\mathsf{P}_{\geqslant r}}$.

- $\mathsf{P}_{=1}\ulcorner \varphi \to \psi \urcorner \to (\mathsf{P}_{=1}\ulcorner \varphi \urcorner \to \mathsf{P}_{=1}\ulcorner \psi \urcorner)$,

- $\mathsf{P}_{=1}\ulcorner \neg\varphi \urcorner \to \neg\mathsf{P}_{=1}\ulcorner \varphi \urcorner$,

- $\forall n \mathsf{P}_{=1}\ulcorner \varphi \urcorner(\overline{n}/x) \to \mathsf{P}_{=1}\ulcorner \forall x \varphi \urcorner$.

Let $\mathsf{IntroNAddPr}$ be the minimal set of sentences of $\mathcal{L}_{\mathsf{P}_{\geqslant r}}$ such that:

- $\mathsf{IntroNAddPr}$ is closed under first order logical consequence.

- If $\mathbb{N}\mathsf{AddPr} \cup \mathsf{PA} \vdash \varphi$ then $\varphi \in \mathsf{IntroNAddPr}$

- Whenever $\varphi \in \mathsf{IntroNAddPr}$, $\mathsf{P}_{=1}\ulcorner \varphi \urcorner \in \mathsf{IntroNAddPr}$.

Theorem 2.4.12 (McGee (1985)). *Let γ denote a sentence such that:*[8]

$$\mathsf{PA} \vdash \gamma \leftrightarrow \neg\forall n \overbrace{\mathsf{P}_{=1}\ulcorner \mathsf{P}_{=1}\ulcorner \ldots \mathsf{P}_{=1}\ulcorner \gamma \urcorner \urcorner \urcorner}^{n+1}.$$

Then

$$\mathbb{N}\mathsf{AddPr} \cup \mathsf{PA} \vdash \gamma.$$

Furthermore, $\mathsf{IntroNAddPr}$ *contains the following sentences:*

---

[7]Though in fact only Robinson arithmetic is needed.

[8]This is an informal ascription of the formula $\neg\forall x(N(x) \to \mathsf{P}_{=1}(g(\mathsf{S}(x), \ulcorner \gamma \urcorner)))$ where g represents the primitive recursive function $G(n, \varphi) = \# \overbrace{\mathsf{P}_{=1}\ulcorner \ldots \mathsf{P}_{=1}\ulcorner \varphi \urcorner \ldots \urcorner}^{n}$.

- γ, *and therefore* $\neg \forall n \overbrace{\mathsf{P}_{=1} \ulcorner \mathsf{P}_{=1} \ulcorner \ldots \mathsf{P}_{=1} \ulcorner \gamma \urcorner \urcorner \urcorner}^{n+1}$

- $\mathsf{P}_{=1} \ulcorner \gamma \urcorner$

- $\mathsf{P}_{=1} \ulcorner \mathsf{P}_{=1} \ulcorner \gamma \urcorner \urcorner$

- $\ldots$

Before proving this, we will explicitly state the important corollary of this theorem:

Corollary 2.4.13. $\mathsf{IntroNAddPr}$ *is ω-inconsistent.*

Proof of Theorem 2.4.12. The following is a derivation using NAddPr and PA:

$$\neg \gamma \to \forall n \in N \overbrace{\mathsf{P}_{=1} \ulcorner \mathsf{P}_{=1} \ulcorner \ldots \mathsf{P}_{=1} \ulcorner \gamma \urcorner \urcorner \urcorner}^{n+1}$$

$$\to \forall n \in N \mathsf{P}_{=1} \ulcorner \overbrace{\mathsf{P}_{=1} \ulcorner \mathsf{P}_{=1} \ulcorner \ldots \mathsf{P}_{=1} \ulcorner \gamma \urcorner \urcorner \urcorner}^{n+1} \urcorner$$

$$\to \mathsf{P}_{=1} \ulcorner \forall n \in N \overbrace{\mathsf{P}_{=1} \ulcorner \mathsf{P}_{=1} \ulcorner \ldots \mathsf{P}_{=1} \ulcorner \gamma \urcorner \urcorner \urcorner}^{n+1} \urcorner$$

$$\to \neg \mathsf{P}_{=1} \ulcorner \neg \forall n \in N \overbrace{\mathsf{P}_{=1} \ulcorner \mathsf{P}_{=1} \ulcorner \ldots \mathsf{P}_{=1} \ulcorner \gamma \urcorner \urcorner \urcorner}^{n+1} \urcorner$$

$$\to \neg \mathsf{P}_{=1} \ulcorner \gamma \urcorner$$

$$\to \neg \forall n \in N \overbrace{\mathsf{P}_{=1} \ulcorner \mathsf{P}_{=1} \ulcorner \ldots \mathsf{P}_{=1} \ulcorner \gamma \urcorner \urcorner \urcorner}^{n+1}$$

$$\to \gamma$$

$$\text{so} \quad \gamma$$

So $\mathsf{NAddPr} \cup \mathsf{PA} \vdash \gamma$ therefore $\gamma \in \mathsf{IntroNAddPr}$ and so also

$$\neg \forall n \in N \overbrace{\mathsf{P}_{=1} \ulcorner \mathsf{P}_{=1} \ulcorner \ldots \mathsf{P}_{=1} \ulcorner \gamma \urcorner \urcorner \urcorner}^{n+1} \in \mathsf{IntroNAddPr}.$$

Therefore $\mathsf{P}_{=1} \ulcorner \gamma \urcorner \in \mathsf{IntroNAddPr}$. Therefore $\mathsf{P}_{=1} \ulcorner \mathsf{P}_{=1} \ulcorner \gamma \urcorner \urcorner \in \mathsf{IntroNAddPr}$. Etc. $\square$

Proof of Theorem 2.4.10. We prove this by using Theorem 2.4.12 and two additional lemmas.

Lemma 2.4.14. *Let $\mathfrak{M}$ be a probabilistic modal structure where m_w is countably additive and each $\mathbf{M}(v)$ is an $\mathbb{N}$-model. If $\mathbf{p} = \Theta_{\mathfrak{M}}(\mathbf{p})$ then*

$$(\mathbf{M}, \mathbf{p})(w) \models \mathsf{NAddPr} \cup \mathsf{PA}$$

Proof. Suppose φ a theorem of Peano arithmetic. Then $\varphi \in \mathsf{Sent}_{\mathcal{L}}$ so $(w, f) \models_{\mathfrak{M}} \varphi \iff \mathbf{M}(w) \models \varphi$. We assumed that for each w, $\mathbf{M}(w)$ interprets the arithmetic vocabulary by the standard model of arithmetic, $\mathbb{N}$, so it must be that $\mathbf{M}(w) \models \varphi$ and so $(\mathbf{M}, \mathbf{p})(w) \models \varphi$.

46

The interesting case for $\mathbb{N}\mathsf{AddPr}$ is that for the quantifier, which is where we use the countable additivity of m_w:

To show $\forall n \mathsf{P}_{=1} \ulcorner \varphi \urcorner (\overline{n}/x) \rightarrow \mathsf{P}_{=1} \ulcorner \forall x \varphi \urcorner$:

$$(\mathbf{M}, \mathbf{p})(w) \models \forall n \mathsf{P}_{=1} \ulcorner \varphi \urcorner (\overline{n}/x)$$

$\implies$ for all n, $(\mathbf{M}, \mathbf{p})(w) \models \mathsf{P}_{=1} \ulcorner \varphi[\overline{n}/x] \urcorner$

$\implies$ for all n, $m_w \{v \mid (\mathbf{M}, \mathbf{p})(v) \models \varphi[\overline{n}/x]\} = 1 \quad \mathbf{p} = \Theta_{\mathfrak{M}}(\mathbf{p})$

$\implies m_w \left(\bigcap \{v \mid (\mathbf{M}, \mathbf{p})(v) \models \varphi[\overline{n}/x]\} \right) = 1 \qquad m_w$ is countably additive

$\implies m_w \{v \mid \text{for all } n, (\mathbf{M}, \mathbf{p})(v) \models \varphi[\overline{n}/x]\} = 1$

$\implies m_w \{v \mid (\mathbf{M}, \mathbf{p})(v) \models \forall x \varphi\} = 1$

$\implies (w, \mathbf{p}) \models \mathsf{P}_{=1} \ulcorner \forall x \varphi \urcorner$ $\qquad\qquad\qquad\qquad\qquad \square$

Lemma 2.4.15. *Let $\mathfrak{M}$ be a probabilistic modal structure where each m_w is countably additive and each $\mathbf{M}(w)$ is an $\mathbb{N}$-model. If $\mathbf{p} = \Theta_{\mathfrak{M}}(\mathbf{p})$, then for each $w \in W$,*

$$(\mathbf{M}, \mathbf{p})(w) \models \mathsf{Intro}\mathbb{N}\mathsf{AddPr},$$

where $\mathsf{Intro}\mathbb{N}\mathsf{AddPr}$ is as defined in Theorem 2.4.12.

Proof. Suppose we have such a $\mathfrak{M}$ and $\mathbf{p} = \Theta_{\mathfrak{M}}(\mathbf{p})$. We will work by induction on the construction of $\mathsf{Intro}\mathbb{N}\mathsf{AddPr}$. By Lemma 2.4.14, if $\mathbb{N}\mathsf{AddPr} \cup \mathsf{PA} \vdash \varphi$, then for each $w \in W$, $(\mathbf{M}, \mathbf{p})(w) \models \varphi$.

$\{\varphi \mid (\mathbf{M}, \mathbf{p})(w) \models \varphi\}$ is closed under first-order logical consequence because $(\mathbf{M}, \mathbf{p})(w)$ is just given by the first order model $(\mathbf{M}(w), \mathbf{p}(w))$.

Now for the inductive step: suppose $\varphi \in \mathsf{Intro}\mathbb{N}\mathsf{AddPr}$. Then we have by the induction hypothesis that for all $v \in W$, $(\mathbf{M}, \mathbf{p})(v) \models \varphi$. Therefore

$$m_w \{v \mid (\mathbf{M}, \mathbf{p})(v) \models \varphi\} = 1$$

and so, since $\mathbf{p} = \Theta_{\mathfrak{M}}(\mathbf{p})$, $(\mathbf{M}, \mathbf{p})(w) \models \mathsf{P}_{=1} \ulcorner \varphi \urcorner$. $\qquad\qquad \square$

This suffices for our result because since $\mathbf{M}(w)$ is a standard model of arithmetic and so $\{\varphi \mid (\mathbf{M}, \mathbf{p})(w) \models \varphi\}$ must be ω-consistent. This therefore contradicts Theorem 2.4.12. So if $\mathfrak{M}$ is as described, then there can be no $\mathbf{p}$ with $\mathbf{p} = \Theta_{\mathfrak{M}}(\mathbf{p})$. $\mathfrak{M}$ therefore does not support a Prob-PW-model. $\qquad \square$

This rules out a large swathe of frames and in particular also all finite frames. However, it is also important to note the limitation of this result: it does not apply to frames that are merely-finitely-additive. We will generally restrict our attention to frames which are finitely-additive and moreover often merely-finitely-additive since we have assumed that the accessibility measure is defined on the powerset algebra. This result therefore does not apply to the frames that we will generally be focusing on in this thesis. However, we do not yet have a result that says that any merely finitely additive frames *do* support a Prob-PW-model.

2.5 The strategy of ruling out probabilistic modal structures because of inconsistencies

In Section 1.4 we showed that the introspection principles in such an expressively rich language (in particular, one including π) is inconsistent with the assumption

of probabilistic coherence. In Caie (2013), Caie takes this as a *prima facie* argument against probabilism. Egan and Elga (2005) consider related cases, more like *Promotion*, and argue that an agent shouldn't believe the equivalence between the sentence and the facts about her degrees of belief. Such a response isn't available in the case of π since the equivalence $\pi \leftrightarrow \neg P_{\geqslant 1/2} \ulcorner \pi \urcorner$ is derivable in arithmetic. A third option is to instead reject introspection.

The result in Section 2.4.2, that no weakly introspective probabilistic modal structure supports a Prob-PW-model, was a statement of this conflict between introspection and probabilism in the framework of probabilistic modal structures. And the option of rejecting introspection in this framework is to reject probabilistic modal structures that do not support a Prob-PW-model. This therefore provides a restriction on the class of admissible, or acceptable, probabilistic modal structures.

However, Theorem 2.4.10 shows that a large class of probabilistic modal structures do not support Prob-PW-models. So if we are to take this approach we would have to also rule out any finite or countably additive structures, or at least those where arithmetic is interpreted using the standard model of arithmetic.[9] This suggests that it is the wrong approach as it is too restrictive. Instead we should *not* reject such structures, so should not reject the possibility that agents are introspective, but should instead account for how we can deal with such structures. This is the approach we will be taking in this thesis. In fact we will provide semantics that differ from the Prob-PW-models and we will show that then the principles expressing introspection should be reformulated with a truth predicate.

2.6 Options for developing a semantics and an overview of Part I

We have seen that there are challenges facing the development of a semantics based on probabilistic modal structures. Since we also want to allow for structures that do not support Prob-PW-models, we need to provide some semantics that works differently from just Prob-PW-models.

Probabilistic modal structures are useful for allowing varying interpretations of probability and the flexibility required for modelling subjective probability so we would like to come up with an alternative semantics that can still work over such structures. So what alterations can we make to the naive definition? The fact that some structures do not admit Prob-PW-models, and other challenges arising from self-referential probabilities, is very much connected to the liar paradox. There has been a lot of work on the liar paradox, developing semantics and theories of truth. We will generalise such semantics and theories of truth to also apply to probability. There has also been work on predicate approaches to necessity, and other all-or-nothing modalities like knowledge, where the semantics for truth have been generalised to apply over possible world structures in the form of Kripke spaces. Notable work in this is in Halbach et al. (2003); Halbach and Welch (2009); Stern (2015b, 2014a,b). Some of the generalisations that we present in this thesis will be similar in technical spirit to the

---

[9]And all introspective probabilistic modal structures, also those involving a non-standard model of arithmetic.

work in those papers. For example, in Halbach et al. (2003), Halbach, Leitgeb and Welch consider the question directly analogous to the question of which probabilistic modal structures support Prob-PW-models.

There are two very influential theories and semantics for truth which we will be considering and generalising in this thesis. The first is a Kripke-style semantics, the origins of which are in Kripke (1975), and the second is a revision theory, conceived of by Herzberger and Gupta and presented in Gupta and Belnap (1993). The generalisations that we consider typically work by also adding the probabilistic modal structure.

In Chapters 3 and 4 we will consider the generalisation of the Kripke-style semantics. A Kripke-style semantics drops certain aspects of classical logic, in particular we may have that neither $\mathsf{T}\ulcorner\lambda\urcorner$ nor $\mathsf{T}\ulcorner\neg\lambda\urcorner$ are satisfied, so classical logic doesn't hold inside the truth predicate. Similarly we will adopt some "non-classical" probabilities. In Chapter 3, the semantics developed is best understood as assigning intervals of probability values to sentences. The intervals assigned can be seen as the appropriate version of probabilities when one adopts a strong Kleene evaluation schemes instead of classical logic. These, for example will not have that $\lambda \vee \neg\lambda$ is assigned probability 1, but that will instead be assigned the unit interval. In Chapter 4 we have *imprecise probabilities* that interpret the probability notion by *sets of* probability functions. Interest in imprecise probabilities has been growing in recent years and this provides and interesting connection between self-referential probabilities and imprecise probabilities, which could also be viewed as an alternative kind of argument for imprecise probabilities.

In developing a Kripke-style semantics one needs to choose a partial evaluation scheme and this choice is what distinguishes Chapter 3 from Chapter 4. The scheme we focus on in Chapter 3 is a strong Kleene evaluation scheme. This is particularly interesting because we can obtain an axiomatisation of this semantics that is complete if one assumes an ω-rule. In Chapter 4 the underlying logic is a supervaluation one. In these imprecise probabilities models we have the nice feature that at fixed points every member of the credal state looks best from some (possibly different) member's perspective.

In Chapter 5 we will consider a revision theory of probability. In this chapter we will consider two main variants, the first interprets the probability notion by considering the relative frequency of the sentences being true and the second is again based on probabilistic modal structures. For the first variant, the probability notion developed should be interpreted as something like semantic probability, see Section 1.2.2. The second variant can be used for subjective probability, or objective chance, or any notion of probability that can be seen to be appropriately modelled by probabilistic modal structures. This second variant is therefore also much more closely related to the semantics of Chapters 3 and 4. In this chapter we will be particularly interested in what happens at the limit stages and will give a definition which provides us with nice models at the limit stage: they will interpret truth with a maximally consistent set of sentences and interpret probability using a function that is probabilistically coherent.

2.7 Conditional probabilities revisited

We discussed in Section 1.7 that the ratio formula may not appropriately capture conditionalisation in the higher-order setting. We can now make this idea more formal by using the probabilistic modal structures to present a suggested way that conditional probabilities should work. Within a probabilistic modal structure, updating or conditionalisation can be understood as learning a partition.[10]

2.7.1 Updating in a probabilistic modal structure

Suppose we start with a probabilistic modal structure $\mathfrak{M}$ modelling some situation. Let's just suppose that there is only one agent that is being modelled.[11] A learning situation where the agent learns the partition Π is formulated by moving to the new probabilistic modal structure $\mathfrak{M}^\Pi$ which modifies $\mathfrak{M}$ just by altering the accessibility relation to:

$$m_w^\Pi(A) := m_w(A \mid S)$$

for $w \in S \in \Pi$. This is always well defined if for every $S \in \Pi$ and $w \in S$, $m_w(S) > 0$. It is ill-defined if an agent learns something where he is probabilistically certain of its negation. Those are cases that we will not consider.

Definition 2.7.1. Let $\mathfrak{M}$ be a probabilistic modal structure and Π a partition of W such that $m_w(S) > 0$ for every $S \in \Pi$ and $w \in S$. Define $\mathfrak{M}^\Pi = (W, m_w^\Pi, \mathbf{M})$ where for $w \in S \in \Pi$:
$$m_w^\Pi(A) := m_w(A \mid S).$$

Example 2.7.2 (Example 1.7.2 ctd.)**.** Suppose we have a setup as in Example 1.7.2, so an agent who is considering the outcome of a toss of a fair coin. We can model her before any learning event by the probabilistic modal structure

Figure 2.6: Agent considering the toss of a fair coin before learning. $\mathfrak{M}$.

Here she just learns the outcome of the coin, so she learns the partition

$$\Pi = \{\{w_{Heads}\}, \{w_{Tails}\}\}$$

So the new probabilistic modal structure is $\mathfrak{M}^\Pi$, presented in Fig. 2.7.

Here is a different example of learning where the agent does not learn exactly which world she is in.

---

[10]Note that we are here focusing on just one agent modelled in a probabilistic modal structure, things become more complicated if multiple agents are considered.

[11]If one considers multiple agents, the learning proposed here characterises a public announcement, available to all agents.

Figure 2.7: Agent considering the toss of a fair coin after learning the outcome of the toss. $\mathfrak{M}^{\Pi}$.

Example 2.7.3. In Example 2.1.2, we modelled an agent where there is an urn with 30 red balls, 30 blue and 30 red, and where a random ball is drawn and the agent is told whether it is yellow or not.

This was modelled by the structure
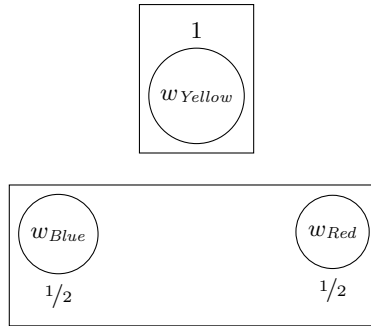
Figure 2.8: Example 2.1.2, $\mathfrak{M}^{\Pi}$.

This probabilistic modal structure can be seen as formed from updating in the suggested way.

Here, the partition that she learns is

$$\Pi = \{\{w_{Yellow}\}, \{w_{Blue}, w_{Red}\}\}$$

which is the information about whether the ball is yellow or not. If the original probabilistic modal structure is given by $\mathfrak{M}$ as in Fig. 2.9, then this new probabilistic modal structure is exactly $\mathfrak{M}^{\Pi}$ according to the definition of learning formulated within a probabilistic modal structure.
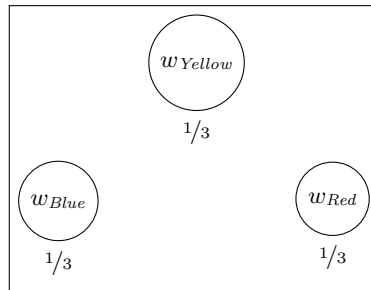
Figure 2.9: Example 2.1.2 before learning, $\mathfrak{M}$.

2.7.2 The ratio formula doesn't capture updating

So using this idea of how updating works, we can now formalise the thought presented in Section 1.7 that Bayes formula does not appropriately capture updating.

Example 2.7.4 (Example 2.7.2 ctd.)**.** We can use $\mathfrak{M}$ and $\mathfrak{M}^{\Pi}$ from our coin tossing case to also calculate probabilities on the sentential level.

Using this structure we can now determine:

$$w_{Heads} \models_{\mathfrak{M}} Heads$$
$$w_{Tails} \models_{\mathfrak{M}} \neg Heads$$
$$w_{Heads} \models_{\mathfrak{M}} \mathbb{P}_{=^1/_2}(Heads)$$
$$w_{Tails} \models_{\mathfrak{M}} \mathbb{P}_{=^1/_2}(Heads)$$
$$w_{Heads} \models_{\mathfrak{M}} \mathbb{P}_{=0}(\mathbb{P}_{=1}(Heads) \mid Heads)$$

Where the last, conditional probability is defined by the ratio formula, and shows that Eq. (1.7) from Section 1.7 can be seen as determining, with the ratio formula, the conditional probability in the pre-learning probabilistic modal structure.

But, the (unconditional) probability in the updated structure, $\mathfrak{M}^{\Pi}$ is given by:

$$w_{Heads} \models_{\mathfrak{M}^{\Pi}} Heads$$
$$w_{Tails} \models_{\mathfrak{M}^{\Pi}} \neg Heads$$
$$w_{Tails} \models_{\mathfrak{M}^{\Pi}} \mathbb{P}_{=1}(Heads)$$
$$w_{Heads} \models_{\mathfrak{M}^{\Pi}} \mathbb{P}_{=1}(\mathbb{P}_{=1}(Heads))$$

which is a formalisation of Eq. (1.6).

And we see here that the conditional probability in the old structure as given by the ratio formula does not equal the probability in the updated structure.

So the ratio formula is not appropriately formalising updating. This doesn't mean that what is expressed with the ratio formula is not useful. It may, for example, appropriately express supposition. This might be the difference already discussed: But this shows that it needs to be further studied and in this thesis we will not discuss conditional probabilities.

The problem is that the worlds where $\mathbb{P}_{=1}(Heads)$ is true changes between $\mathfrak{M}$ and $\mathfrak{M}^{\Pi}$.

$$\{v \in W \mid v \models_{\mathfrak{M}} \mathbb{P}_{=1}(Heads)\} \neq \{v \mid v \models_{\mathfrak{M}^{\Pi}} \mathbb{P}_{=1}(Heads)\}.$$

For calculating the ratio formula we use the old interpretation of $\mathbb{P}_{=1}(Heads)$, whereas in the updated structure we look at the updated interpretation of $\mathbb{P}_{=1}(Heads)$.

We can get updating appropriately modelled in the object language by defining $\mathbb{P}_{=?}(\cdot \mid\mid \cdot)$ to be what one actually gets from updating.

Definition 2.7.5. Consider a language $\mathcal{L}_{\mathbb{P}_{\geqslant r}(\cdot \mid\mid \cdot)}$ which adds a binary operator, so if φ and ψ are sentences then so are $\mathbb{P}_{\geqslant r}(\varphi \mid\mid \psi)$ for any r rational.

For ease of writing, we make the following definitions (which will only be used in this section).

Definition 2.7.6. Define
$$[\varphi]_{\mathfrak{M}} := \{v \mid v \models_{\mathfrak{M}} \varphi\}.$$

Let
$$\mathfrak{M}^{\psi} := \mathfrak{M}^{\{[\psi]_{\mathfrak{M}}, W \setminus [\psi]_{\mathfrak{M}}\}}.$$

So $\mathfrak{M}^{\psi}$ denotes the probabilistic modal structure after the agent has learned, in each world, whether or not φ is the case.

We can now give the semantics for this language, with the important clause:

Definition 2.7.7. If $w \models_{\mathfrak{M}} \psi$, we define:
$$w \models_{\mathfrak{M}} \mathbb{P}_{\geqslant r}(\varphi \,\|\, \psi) \iff m_w([\varphi]_{\mathfrak{M}^{\psi}} \mid [\psi]_{\mathfrak{M}}) \geqslant r$$

This now does express what the probability would be if the agent were to learn ψ by virtue of the following result.

Proposition 2.7.8. *If $w \models_{\mathfrak{M}} \psi$,*
$$w \models_{\mathfrak{M}} \mathbb{P}_{\geqslant r}(\varphi \,\|\, \psi) \iff w \models_{\mathfrak{M}^{\psi}} \mathbb{P}_{\geqslant r} \varphi$$

2.7.3 Analysis of this language

One should then consider properties of this new conditional probability. We can take this as primitive and drop the unconditional probability, defining:

Proposition 2.7.9.
$$w \models_{\mathfrak{M}} \mathbb{P}_{\geqslant r} \varphi \iff w \models_{\mathfrak{M}} \mathbb{P}_{\geqslant r}(\varphi \,\|\, \top)$$

But we do not have any rule like the ratio formula which would allow us to just take unconditional probability as primitive and define conditional probability from it.

We might also ask: when does the ratio formula appropriately represent updating, i.e. when is $(\cdot \mid \cdot)$ the same as $(\cdot \,\|\, \cdot)$? This will be the case if
$$m_w([\varphi]_{\mathfrak{M}^{\psi}} \mid [\psi]_{\mathfrak{M}}) = m_w([\varphi]_{\mathfrak{M}} \mid [\psi]_{\mathfrak{M}})$$

when $w \in [\psi]_{\mathfrak{M}}$. If φ doesn't talk about probabilities, then $[\varphi]_{\mathfrak{M}} = [\varphi]_{\mathfrak{M}^{\psi}}$ and so then the equivalence will hold.

Proposition 2.7.10. *If φ in $\mathcal{L}$, then $[\varphi]_{\mathfrak{M}} = [\varphi]_{\mathfrak{M}^{\psi}}$ and so for any $\psi \in \mathcal{L}_{\mathbb{P}_{\geqslant r}}$*
$$w \models_{\mathfrak{M}} \mathbb{P}_{\geqslant r}(\varphi \mid \psi) \iff w \models_{\mathfrak{M}} \mathbb{P}_{\geqslant r}(\varphi \,\|\, \psi)$$

So it is really the higher order probabilities in the target sentence which cause the problem. Just having higher order probabilities in what is being learned does not affect the appropriateness of the ratio formula for representing updating. For example:

Example 2.7.11.

$$w \models_{\mathfrak{M}} \mathbb{P}_{\geqslant r}(Heads \mid \mathbb{P}_{=1/2}(Heads)) \iff w \models_{\mathfrak{M}} \mathbb{P}_{\geqslant r}(Heads \,\|\, \mathbb{P}_{=1/2}(Heads))$$

It should be investigated how this proposal connects to Romeijn (2012) and to work in dynamic epistemic logic. But that lies outside the scope of this thesis. For the rest of this thesis only unconditional probability will be considered.

Chapter 3

A Kripkean Theory

3.1 Introduction

In this chapter we will develop a Kripke-style theory of truth. This will generalise a very influential theory of truth that originates in a paper by Saul Kripke (1975). Kripke's theory of truth was developed to account for a languages with type-free truth predicates and can therefore express the liar sentence. In his paper Kripke constructs an extension of the truth predicate by formalising the procedure of evaluating a sentence. He uses three-valued evaluation schemes to build up this extension, but the extension of the truth predicate can also be used within classical logic to give a classical model of the language with a truth predicate. In this semantics one will have that, for all sentences. For example, for the liar sentence, neither $\mathsf{T}\ulcorner\varphi\urcorner$ nor $\mathsf{T}\ulcorner\neg\varphi\urcorner$ are satisfied.

In this chapter we shall present a generalisation of this semantics to also account for probability predicates. The final semantics we propose will not determine particular point valued probabilities for some sentences. For example we might have that neither $\mathsf{P}_{>}(\ulcorner\varphi\urcorner,\ulcorner 0\urcorner)$ nor $\mathsf{P}_{<}(\ulcorner\varphi\urcorner,\ulcorner 1\urcorner)$, but the only information that we have about the probability of φ is that both $\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner 0\urcorner)$ and $\mathsf{P}_{\leqslant}(\ulcorner\varphi\urcorner,\ulcorner 1\urcorner)$ are satisfied. In that case we would say that φ is assigned the interval of values $[0,1]$.

Our generalisation follows ideas from Halbach and Welch (2009) where Halbach and Welch develop a semantics for necessity, conceived of as a predicate, by applying Kripke's construction to "possible world" structures in the form of Kripke models from modal logic. We will use probabilistic modal structures to provide the background structure for our construction. This therefore allows one to use the technical advantages of these structures which might have been thought to only be available when the probability notion is conceived of as an operator (see Halbach et al., 2003).

The language we will work with will have a truth predicate and a a probability predicate. This language will formalise the probability notion as a predicate that applies to the codes of sentences and rational numbers. We will have a sentence like "$\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)$" whose intended interpretation is "The probability of φ is $\geqslant r$".

Outline of the chapter

The chapter is structured as follows.

In Section 3.2, we will motivate and present our suggested semantics. As suggested, this will generalise Kripke's theory of truth by applying it over probabilistic modal structures. The general strategy follows Halbach and Welch (2009).

In Section 3.4 we give some observations regarding the developed semantics. In Stern (2014a,b), Stern argues that when stating principles about necessity, the job of quotation and disquotation should be done by a truth predicate. We argue for the same thing here: we argue that principles such as the introspection principles are properly expressed by using the truth predicate. In our language the introspection principles will then be written as:

$$\mathsf{T}\ulcorner\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)\urcorner \implies \mathsf{P}_{=}(\ulcorner\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)\urcorner,\ulcorner 1\urcorner)$$
$$\mathsf{T}\ulcorner\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)\urcorner \implies \mathsf{P}_{=}(\ulcorner\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)\urcorner,\ulcorner 1\urcorner)$$

This allows one to avoid inconsistency and is well-motivated in this semantic construction. In Section 3.4 we also consider countable additivity and show that if the underlying probabilistic modal structure has countably additive probability measures, then the resulting semantics will satisfy the version of $\mathbb{N}$-additivity that is appropriate in our framework. $\mathbb{N}$-additivity says:

$$\mathrm{p}\left(\exists x\varphi(x)\right) = \lim_{n\longrightarrow\infty}\mathrm{p}\left(\varphi(\overline{0})\vee\varphi(\overline{1})\vee\ldots\vee\varphi(\overline{n})\right).$$

This is interesting because $\mathbb{N}$-additivity has proved challenging in previous work on self-referential probabilities. Both at the final stage of Leitgeb's construction and in the construction by Christiano et al., there is a formula $\varphi(x)$ such that $P(\exists x\varphi(x)) = 0$ but for each n $P(\varphi(0)\vee\ldots\vee\varphi(\overline{n})) = 1$. This shows that they badly fail $\mathbb{N}$-additivity.

In Section 3.5 we shall give an axiomatic theory that is intended to capture the semantics. Such a theory is important because it allows one to reason about the semantics. As was discussed in Aumann (1999), when one gives a possible worlds framework to formalise a game theory context the question arises of what the players know about the framework itself and this question is best answered by providing a corresponding syntactic approach. Our theory is complete in the presence of the ω-rule, which allows one to conclude $\forall x\varphi(x)$ from all the instances of $\varphi(\overline{n})$. This is needed to fix the standard model of arithmetic. To show the completeness when the ω-rule is present we construct a canonical model. This axiomatisation is substantially new research.

Finally, we finish the paper with some conclusions in Section 3.6.

3.2 A Kripke-style semantics

3.2.1 Setup: language and notation

We will use much of the setup from Section 1.6. The syntax of the language which we focus on in this chapter with will be as follows:

Setup 3 (for Chapter 3). *Let $\mathcal{L}$ be some language extending L_{PA}. We allow for the addition of contingent vocabulary but for technical ease we shall only*

allow contingent relation symbols (and propositional variables) and not function symbols or constants.[1] We also only allow for a countable number of contingent vocabulary symbols in order for our language to remain countable so we can use arithmetic for Gödel coding and for the completeness proof to work.

Let $\mathcal{L}_{\mathsf{P}_{\geqslant},\mathsf{T}}$ extend this language by adding a unary predicate T and a binary predicate $\mathsf{P}_{\geqslant}$ (see Section 1.6.3).

We could consider languages with multiple probability notions, then we would add the binary predicate $\mathsf{P}_{\geqslant}^{\mathsf{A}}$ for each notion of probability, or agent A, but our constructions will immediately generalise to the multiple probability languages so we just focus on the language with one probability notion. We have included the truth predicate since it is easy to extend the definition of the semantics to deal with truth as well as probability, and it is nice to see that the construction can give a joint theory of truth and probability. Additionally, we shall rely on the truth predicate for our later axiomatisation and for expressing principles such as introspection.

We will assume some Gödel coding of expressions and rational numbers, and the corresponding notion, for example $\dot{\neg}$ and $1\dot{-}$, as set up in Definitions 1.6.2 and 1.6.6.

We now introduce the other probability predicates, which we use as abbreviations.

Definition 3.2.1. Define for terms t and s the following abbreviations:

- $\mathsf{P}_{>}(t,s) := \exists x \succ s(\mathsf{P}_{\geqslant}(t,x))$

- $\mathsf{P}_{\leqslant}(t,s) := \mathsf{P}_{\geqslant}(\dot{\neg}t, 1\dot{-}s)$

- $\mathsf{P}_{<}(t,s) := \mathsf{P}_{>}(\dot{\neg}t, 1\dot{-}s)$

- $\mathsf{P}_{=}(t,s) := \mathsf{P}_{\geqslant}(t,s) \wedge \mathsf{P}_{\leqslant}(t,s)$

In a model that interprets the arithmetic vocabulary by the standard model of arithmetic we will have that $\mathsf{P}_{>}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)$ holds if and only if there is some $q > r$ such that $\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner q\urcorner)$ holds.

3.2.2 The construction of the semantics

We will now move to developing our semantics.

Kripke's construction (from Kripke, 1975) is motivated by the idea that one should consider the process of evaluating a sentence to determine which sentences can unproblematically be given a truth value.

To evaluate the sentence $\mathsf{T}\ulcorner 0 = 0\urcorner$ one first has to evaluate the sentence $0 = 0$. Since $0 = 0$ does not mention the concept of truth it can easily be evaluated so $\mathsf{T}\ulcorner 0 = 0\urcorner$ can then also be evaluated. Kripke formalises this process of evaluating sentences. We shall say evaluated positively (and evaluated negatively) instead of evaluated as true (and evaluated as false) to make it clear that this is happening at the meta-level.

To evaluate $\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)$ we first need to evaluate φ not only in the actual state of affairs but also in some other states of affairs. We therefore base our

---

[1] This restriction could be dropped, but then $t^{\mathbb{N}}$ (as will be defined in Definition 1.6.2) would only be defined for terms of L_{PA} instead of arbitrary terms of $\mathcal{L}$ and this would then just complicate the presentation of the material.

construction on structures with multiple "possible worlds" and we evaluate the sentences at all the worlds. We will assume that each world has a "degree of accessibility" relation to the other worlds. This will be used to give us the interpretation of $P_{\geqslant}$.

To do this we will use the *probabilistic modal structures* as introduced in Definition 2.1.1. Remember these were given by a frame and valuation. A frame consists of a collection of "possible worlds", W, and a collection of probability measures m_w over W. For simplification we are assuming in this chapter that we only have one agent, so we only have a single m_w for each world w. The valuation assigns to each world a model of the language without probability or truth, $\mathbf{M}(w)$. For this chapter we will assume that each $\mathbf{M}(w)$ has the natural numbers as a domain and interprets the arithmetic vocabulary in the standard way,[2] so these essentially just give interpretations to the empirical predicates. As before, we call such models $\mathbb{N}$-models.

We now move to motivating our construction of the extension of the probability predicate. At the first stage we use $\mathbf{M}$ to see that *Blue* is evaluated positively in w_{Blue} and negatively in the other worlds. So using the frame we see that at the second stage we should now evaluate $P_{\geqslant}(\ulcorner Blue \urcorner, \ulcorner 1/2 \urcorner)$ positively in w_{Blue} and w_{Red} and negatively in w_{Yellow}.

To formalise the evaluation procedure we need to record how the sentences have been evaluated at each world. We do this by using an *evaluation function* that records the codes of the sentences that are evaluated positively. In doing this we only focus on those sentences that are evaluated positively and see that φ is evaluated negatively if and only if $\neg\varphi$ is evaluated positively.

Definition 3.2.2. An *evaluation function*, f, assigns to each world, w, a set $f(w) \subseteq \mathbb{N}$.
   If $\#\varphi \in f(w)$, we say that f *evaluates φ positively at w*.

This bears a close relationship to a prob-eval function. An evaluation function can be seen as a considered interpretation of the truth predicate at each world, whereas a prob-eval function gave a considered interpretation of the probability predicates. Here we are interested in the connection between truth and probability and it will simplify matters to just consider the extension truth predicate. We can then determine the extension of the probability predicates from this by using the underlying probabilistic modal structure. In doing this we are following the setup from Stern (2015a).

We can now proceed to give a formal analysis of the evaluation procedure. We do this by developing a definition of $\Theta(f)$, which is the evaluation function given by another step of reasoning. So if f gives the codes of the sentences that we have so far evaluated positively, then $\Theta(f)$ gives the codes of the sentences that one can evaluate positively at the next stage.

At the zero-th stage one often starts without having evaluated any sentence either way. This can be given by an evaluation function f_0 with $f_0(w) = \varnothing$ for all w.

A sentence that does not involve truth or probability can be evaluated positively or negatively by just considering $\mathbf{M}(w)$. So we define:

---

[2]This restriction of the domain allows us to have a name for each member of the domain and therefore makes the presentation easier since we can then give the semantics without mentioning open formulas and variable assignments. This restriction also helps for the axiomatisation.

- For φ a sentence of $\mathcal{L}$, $\#\varphi \in \Theta(f)(w) \iff \mathbf{M}(w) \models \varphi$

- For φ a sentence of $\mathcal{L}$, $\#\neg\varphi \in \Theta(f)(w) \iff \mathbf{M}(w) \not\models \varphi$

This will give the correct evaluations to the sentences of $\mathcal{L}$, for example $\#0 = 0 \in \Theta(f)(w)$ and $\#\neg 0 = 1 \in \Theta(f)(w)$.

To evaluate a sentence $\mathsf{T}\ulcorner\varphi\urcorner$ we first evaluate φ. If φ was evaluated positively then we can now evaluate $\mathsf{T}\ulcorner\varphi\urcorner$ positively, and similarly if it was evaluated negatively. However, if φ was not evaluated either way then we still do not evaluate $\mathsf{T}\ulcorner\varphi\urcorner$ either way. This is described by the clauses:

- $\#\mathsf{T}\ulcorner\varphi\urcorner \in \Theta(f)(w) \iff \#\varphi \in f(w)$

- $\#\neg\mathsf{T}\ulcorner\varphi\urcorner \in \Theta(f)(w) \iff \#\neg\varphi \in f(w)$

For example we get that $\mathsf{T}\ulcorner 0 = 0\urcorner \in \Theta(\Theta(f))(w)$ and $\neg\mathsf{T}\ulcorner 0 = 1\urcorner \in \Theta(\Theta(f))(w)$.

To describe the cases for probability we consider the fragment of a probabilistic modal frame that is pictured in Fig. 3.1. We consider how one should evaluate $\mathsf{P}_{\geqslant}(\ulcorner\psi\urcorner, \ulcorner r\urcorner)$ for different values of r.
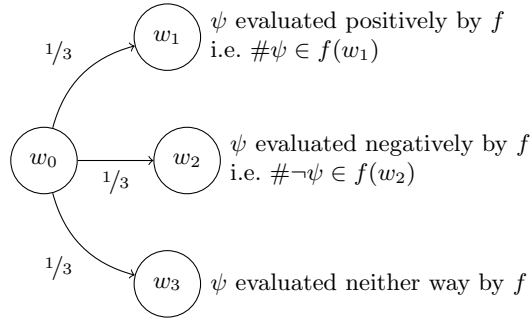
Figure 3.1: A fragment of a probabilistic modal structure representing the information required to evaluate $\mathsf{P}_{\geqslant}(\ulcorner\psi\urcorner, \ulcorner r\urcorner)$ in $\Theta(f)(w_0)$.

$\mathsf{P}_{\geqslant}(\ulcorner\psi\urcorner, \ulcorner 0.3\urcorner)$ will be evaluated positively by $\Theta(f)$ because the measure of the worlds where ψ is evaluated positively is $\frac{1}{3} = 0.333\ldots$, which is larger than 0.3.[3] $\mathsf{P}_{\geqslant}(\ulcorner\psi\urcorner, \ulcorner 0.7\urcorner)$ will be evaluated negatively by $\Theta(f)$ because however ψ will be evaluated in w_3 there are be too many worlds where ψ is already evaluated negatively for the measure of the worlds where it is evaluated positively to become larger than 0.7, while the evaluation function remains consistent this measure could at most become $0.666\ldots = 1 - m_w\{v \mid \#\neg\psi \notin f(v)\}$. We evaluate $\mathsf{P}_{\geqslant}(\ulcorner\psi\urcorner, \ulcorner 0.5\urcorner)$ neither way because if ψ was to become evaluated in w_3 the measure of the worlds where ψ is evaluated positively would become either $0.333\ldots$ or $0.666\ldots$ so we need to retain the flexibility that $\mathsf{P}_{\geqslant}(\ulcorner\psi\urcorner, \ulcorner 0.5\urcorner)$ can later be evaluated either positively or negatively depending on how ψ is evaluated at w_3.

We therefore give the definition

- $\#\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner) \in \Theta(f)(w) \iff m_w\{v \mid \#\varphi \in f(v)\} \geqslant r$

---

[3]One should really say "the measure of *the set of* the worlds where ψ is evaluated positively", but that would be cumbersome.

- $\#\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner) \in \Theta(f)(w) \iff m_w\{v \mid \#\neg\varphi \in f(v)\} > 1 - r$

Consistent evaluation functions are ones where no sentence and its negation both appear in $f(w)$ for any w. For a more precise definition see Definition 3.2.8. These evaluation functions are of particular interest, and for these we have:[4]

$$\#\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner) \in \Theta(f)(w)$$
$$\iff \text{ for all } g \text{ consistent extending } f, \ m_w\{v \mid \#\varphi \in g(v)\} \not\geqslant r$$

In this example we saw that the probability of ψ is given by a range. This is described pictorially in Fig. 3.2.

$\# \cdot \in \Theta(f)(w):$
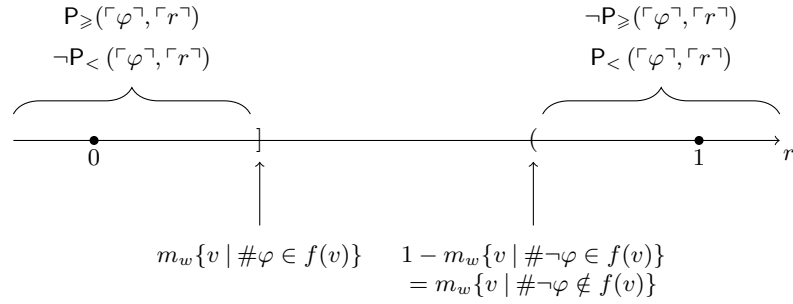
Figure 3.2: How $\Theta(f)(w)$ evaluates the probability of φ.

We lastly need to give the definitions for the connectives and quantifiers. For example we need to say how $\varphi \vee \neg\varphi$ should be evaluated if φ is itself evaluated neither way. For this we directly use the strong Kleene three valued evaluation scheme, which is the scheme that Kripke focused on and there has been a lot of work following him in this. This scheme has that $\#\varphi \vee \psi \in \Theta(f)(w) \iff \#\varphi \in \Theta(f)(w)$ or $\#\psi \in \Theta(f)(w)$, so if φ is evaluated neither way then $\varphi \vee \neg\varphi$ will also be evaluated neither way. The advantage of this scheme over, for example, one

---

[4] This result shows that the semantics developed in (Caie, 2011, Section 4.2.1) is a special case of ours.

This observation holds because:

$$\#\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner) \in \Theta(f)(w)$$
$$\iff m_w\{v \mid \#\neg\varphi \in f(v)\} > 1 - r$$
$$\iff m_w\{v \mid \#\neg\varphi \notin f(v)\} < r$$
$$\iff \text{For all } g \text{ consistent extending } f, \ m_w\{v \mid \#\varphi \in g(v)\} < r$$

The left-to-right of this last step works by: if g is consistent extending f then

$$\{v \mid \#\neg\varphi \notin f(v)\} \supseteq \{v \mid \#\neg\varphi \notin g(v)\} \qquad \text{as } g \text{ extends } f,$$
$$\supseteq \{v \mid \#\varphi \in g(v)\} \qquad \text{as } g \text{ is consistent.}$$

The right-to-left works by constructing a g consistent extending f with

$$\#\varphi \in g(v) \iff \#\neg\varphi \notin f(v),$$
$$\#\neg\varphi \in g(v) \iff \#\neg\varphi \in f(v).$$

based on supervaluational logic is that it is truth functional so the evaluation of $\varphi \vee \psi$ depends only on how φ and ψ have been evaluated.

This fully defines $\Theta(f)$. We only used the question of whether φ can now be evaluated positively, i.e. if $\varphi \in \Theta(f)$, as motivating the definition. We formally understand it as a definition of a three valued semantics $(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}}$ and we will later define $(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \varphi \iff \#\varphi \in \Theta(f)$. This is common when working with Kripke's theory of truth. We sum up our discussion in the formal definition of $(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}}$.

Definition 3.2.3. For $\mathfrak{M}$ a probabilistic modal structure, $w \in W$ and f an evaluation function, define $(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}}$ by induction on the positive complexity of the formula as follows.

- $(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \varphi \iff \mathbf{M}(w) \models \varphi$ for φ an atomic sentence of $\mathcal{L}$

- $(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\varphi \iff \mathbf{M}(w) \not\models \varphi$ for φ an atomic sentence of $\mathcal{L}$

- $(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{T}t \iff t^{\mathbb{N}} \in f(w)$ and $t^{\mathbb{N}} \in \mathsf{Sent}_{\mathsf{P}_{\geqslant},\mathsf{T}}$

- $(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\mathsf{T}t \iff \underset{\cdot}{\neg}t^{\mathbb{N}} \in f(w)$ or $t^{\mathbb{N}} \notin \mathsf{Sent}_{\mathsf{P}_{\geqslant},\mathsf{T}}$

- $(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_{\geqslant}(t, s) \iff m_w\{v \mid t^{\mathbb{N}} \in f(v)\} \geqslant \mathsf{rat}(s^{\mathbb{N}})$ and $s^{\mathbb{N}} \in \mathsf{Rat}$

- $(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\mathsf{P}_{\geqslant}(t, s) \iff m_w\{v \mid \underset{\cdot}{\neg}t^{\mathbb{N}} \in f(v)\} > 1 - \mathsf{rat}(s^{\mathbb{N}})^5$ or $s^{\mathbb{N}} \notin \mathsf{Rat}$

- $(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\neg\varphi \iff (w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \varphi$

- $(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \varphi \vee \psi \iff (w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \varphi$ or $(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \psi$

- $(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg(\varphi \vee \psi) \iff (w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\varphi$ and $(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\psi$

- $(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \exists x \varphi(x) \iff (w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \varphi[\overline{n}/x]$ for some $n \in \mathbb{N}$.

- $(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\exists x \varphi(x) \iff (w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\varphi[\overline{n}/x]$ for all $n \in \mathbb{N}$

The only difference to the standard definition is the addition of the clauses for probability.

As a consequence of our definition we obtain the following results for the other probability variants.

Proposition 3.2.4. *For any probabilistic modal structure $\mathfrak{M}$, evaluation function f and world w the following hold.*

$$(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_{>}(\ulcorner\varphi\urcorner, \ulcorner r \urcorner) \iff m_w\{v \mid \#\varphi \in f(v)\} > r$$
$$(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_{\leqslant}(\ulcorner\varphi\urcorner, \ulcorner r \urcorner) \iff m_w\{v \mid \#\neg\varphi \notin f(v)\} \leqslant r$$
$$(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_{<}(\ulcorner\varphi\urcorner, \ulcorner r \urcorner) \iff m_w\{v \mid \#\neg\varphi \notin f(v)\} < r$$
$$(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_{=}(\ulcorner\varphi\urcorner, \ulcorner r \urcorner) \iff \begin{array}{c} m_w\{v \mid \#\varphi \in f(v)\} \geqslant r \\ \text{and } m_w\{v \mid \#\neg\varphi \notin f(v)\} \leqslant r \end{array}$$

---

[5]Which is $\iff m_w\{v \mid \underset{\cdot}{\neg}t^{\mathbb{N}} \notin f(v)\} < \mathsf{rat}(s^{\mathbb{N}})$

More concisely: if f is consistent (see Definition 3.2.8) we have: for $\rhd \in \{\geqslant$, $>, \leqslant, <, =\}$,

$$(w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{\rhd}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner) \iff \begin{array}{c} m_w\{v \mid \#\varphi \in f(v)\} \rhd r \\ \text{and } m_w\{v \mid \#\neg\varphi \notin f(v)\} \rhd r \end{array}$$

To see the equivalence between this and the above, we observe that for f consistent, $m_w\{v \mid \#\varphi \in f(v)\} \leqslant m_w\{v \mid \#\neg\varphi \notin f(v)\}$.

We can also observe facts about the negated notions as follows: Suppose $\#\varphi \in f(v) \iff \#\neg\neg\varphi \in f(v),$[6] then:

$$(w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\mathsf{P}_{\geqslant}(t, s) \iff (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{<}(t, s), \text{ or } s^{\mathbb{N}} \notin \mathsf{Rat}$$

$$(w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\mathsf{P}_{>}(t, s) \iff (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{\leqslant}(t, s), \text{ or } s^{\mathbb{N}} \notin \mathsf{Rat}$$

$$(w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\mathsf{P}_{\leqslant}(t, s) \iff (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{>}(t, s), \text{ or } s^{\mathbb{N}} \notin \mathsf{Rat}$$

$$(w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\mathsf{P}_{<}(t, s) \iff (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{\geqslant}(t, s), \text{ or } s^{\mathbb{N}} \notin \mathsf{Rat}$$

$$(w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\mathsf{P}_{=}(t, s) \iff (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{>}(t, s) \vee \mathsf{P}_{<}(t, s), \text{ or } s^{\mathbb{N}} \notin \mathsf{Rat}$$

we also have:

$$(w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\mathsf{T}t \iff (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{T}\dot{\neg}t \text{ or } t^{\mathbb{N}} \notin \mathsf{Sent}_{\mathsf{P}_{\geqslant}, \mathsf{T}}$$

Proof. Remember the definition of these from Definition 3.2.1:

- $\mathsf{P}_{>}(t, s) := \exists x \succ s(\mathsf{P}_{\geqslant}(t, x))$
- $\mathsf{P}_{\leqslant}(t, s) := \mathsf{P}_{\geqslant}(\dot{\neg}t, 1\dot{-}s)$
- $\mathsf{P}_{<}(t, s) := \mathsf{P}_{>}(\dot{\neg}t, 1\dot{-}s)$
- $\mathsf{P}_{=}(t, s) := \mathsf{P}_{\geqslant}(t, s) \wedge \mathsf{P}_{\leqslant}(t, s)$

The only interesting case for both of these is the case for $\mathsf{P}_{>}$. However we will write out all the proofs because the results will be used in the proof of soundness and completeness later.

$$(w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{>}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner) \iff (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \exists x \succ \ulcorner r\urcorner(\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, x))$$
$$\iff \exists n, (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} n \succ \ulcorner r\urcorner \wedge \mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, n)$$
$$\iff \exists q > r, (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner q\urcorner)$$
$$\iff \exists q > r, m_w\{v \mid \#\varphi \in f(v)\} \geqslant q$$

which, by the density of $\mathbb{Q}$ in $\mathbb{R}$, is:

$$\iff m_w\{v \mid \#\varphi \in f(v)\} > r$$

$$(w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{\leqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner) \iff (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{\geqslant}(\ulcorner\neg\varphi\urcorner, \ulcorner 1-r\urcorner)$$
$$\iff m_w\{v \mid \#\neg\varphi \in f(v)\} \geqslant 1 - r$$
$$\iff m_w\{v \mid \#\neg\varphi \notin f(v)\} \leqslant r$$

---

[6]Which will be the case after one application of Θ. This is required for the results for $\neg\mathsf{P}_{\leqslant}$, $\neg\mathsf{P}_{<}$ and $\neg\mathsf{P}_{=}$.

$\mathsf{P}_<$ works exactly analogous, so lastly:

$$(w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_=(\ulcorner\varphi\urcorner, \ulcorner r\urcorner) \iff (w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner) \wedge \mathsf{P}_{\leqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)$$

$$\iff \begin{array}{c} m_w\{v \mid \#\varphi \in f(v)\} \geqslant r \\ \text{and } m_w\{v \mid \#\neg\varphi \notin f(v)\} \leqslant r \end{array}$$

For the results with negations we use the fact that $m_w\{v \mid \#\neg\varphi \in f(v)\} = 1 - m_w\{v \mid \#\neg\varphi \notin f(v)\}$.

$$(w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner) \iff m_w\{v \mid \#\neg\varphi \notin f(v)\} < r$$
$$\iff (w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_<(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)$$

$$(w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\mathsf{P}_>(\ulcorner\varphi\urcorner, \ulcorner r\urcorner) \iff (w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\exists x \succ \ulcorner r\urcorner(\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, x))$$
$$\iff \forall q > r, (w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner q\urcorner)$$
$$\iff \forall q > r, m_w\{v \mid \#\neg\varphi \notin f(v)\} \leqslant q$$
$$\iff m_w\{v \mid \#\varphi \in f(v)\} \leqslant r$$

$$(w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\mathsf{P}_{\leqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner) \iff m_w\{v \mid \#\neg\varphi \in f(v)\} > 1 - r$$
$$\iff (w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_>(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)$$

$\mathsf{P}_<$ works exactly analogous, so lastly:

$$(w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\mathsf{P}_=(\ulcorner\varphi\urcorner, \ulcorner r\urcorner) \iff \begin{array}{c} (w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner) \\ \text{or } (w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\mathsf{P}_{\leqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner) \end{array}$$
$$\iff (w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_<(\ulcorner\varphi\urcorner, \ulcorner r\urcorner) \vee \mathsf{P}_>(\ulcorner\varphi\urcorner, \ulcorner r\urcorner) \; \square$$

This shows that our choice of definitions of the derivative probability notions was a good one and that we labelled the diagram of probability ranges as in Fig. 3.2 appropriately.

We now give the definition of Θ in terms of $(w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}}$.

Definition 3.2.5. Define $\Theta_{\mathfrak{M}}$ a function from evaluation functions to evaluation functions by

$$\Theta_{\mathfrak{M}}(f)(w) := \{\#\varphi \mid (w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \varphi\}$$

When $\mathfrak{M}$ is clear from context we will drop reference to it.

We now consider an example of how this works for the "unproblematic" sentences.

Example 3.2.6. Consider again the example in Example 2.1.2 where there is an urn with 30 red, 30 yellow and 30 blue balls and where we are modelling an agent's beliefs after a ball is picked from an urn and the agent is told whether it's yellow or not.

Take any f. Observe that:

$$(w_{Blue}, f) \models_{\mathfrak{M}}^{SKP} Blue \text{ and } (w_{Red}, f) \models_{\mathfrak{M}}^{SKP} \neg Blue$$

so:

$$\#Blue \in \Theta(f)(w_{Blue}) \text{ and } \#\neg Blue \in \Theta(f)(w_{Red}).$$

Therefore:

$$(w_{Blue}, \Theta(f)) \models^{SKP}_{\mathfrak{M}} \mathsf{P}_= (\ulcorner Blue \urcorner, \ulcorner 1/2 \urcorner) \text{ and similarly for } w_{Red}.\text{[7]}$$

so:

$$\#\mathsf{P}_= (\ulcorner Blue \urcorner, \ulcorner 1/2 \urcorner) \in \Theta(\Theta(f))(w_{Blue}) \text{ and similarly for } w_{Red}.$$

Then by similar reasoning:

$$(w_{Blue}, \Theta(\Theta(f))) \models^{SKP}_{\mathfrak{M}} \mathsf{P}_= (\mathsf{P}_= (\ulcorner Blue \urcorner, \ulcorner 1/2 \urcorner), \ulcorner 1 \urcorner)$$

so:

$$\#\mathsf{P}_= (\mathsf{P}_= (\ulcorner Blue \urcorner, \ulcorner 1/2 \urcorner), \ulcorner 1 \urcorner) \in \Theta(\Theta(\Theta(f)))(w_{Blue}).$$

These sentences can be seen as translations of sentences from the operator language. Such sentences will be given point-valued probabilities and be evaluated positively or negatively by some $\Theta(\Theta(\dots \Theta(f) \dots))$. This is described formally in Section 3.3.1.

If one starts with each $f(w) = \varnothing$ and iteratively applies Θ, then Θ will only give evaluations to sentences that were previously evaluated neither way, it will not *change* the evaluation of a sentence. This is because Θ is monotone.

Lemma 3.2.7 (Θ is monotone)**.** *If for all w $f(w) \subseteq g(w)$, then also for all w $\Theta(f)(w) \subseteq \Theta(g)(w)$.*

Proof. Take some evaluation functions f and g such that $f(w) \subseteq g(w)$ for all w. It suffices to prove that if $(w, f) \models^{SKP}_{\mathfrak{M}} \varphi$ then $(w, g) \models^{SKP}_{\mathfrak{M}} \varphi$. This can be done by induction on the positive complexity of φ. $\square$

This is partly due to our observation in Proposition 3.2.4 that the negated notions in fact express positive facts.

This fact ensures that there are fixed points of the operator Θ, i.e., evaluation functions f with $f = \Theta(f)$. These are evaluation functions where the process of evaluation doesn't lead to any new "information".

Definition 3.2.8. f is called a *fixed point evaluation function* if $\Theta(f) = f$.

f is called a *consistent evaluation function* if for each $w \in W$ and $n \in \mathbb{N}$, it is not the case that $n \in f(w)$ and $\dot{\neg} n \in f(w)$.

Corollary 3.2.9 (Θ has fixed points)**.** *For every $\mathfrak{M}$ there is some consistent fixed point evaluation function f.*

Proof. Start with $f_0(w) = \varnothing$ for all w. Let $\Theta^\alpha(f)$ denote the iteration of Θ α-many times to f. Construe each evaluation function as a subset of $W \times \mathbb{N}$. Then by Lemma 3.2.7, $\Theta^\alpha(f_0) \subseteq \Theta^{\alpha+1}(f_0)$. So by cardinality considerations, there must be some β where $\Theta^\beta(f_0) = \Theta^{\beta+1}(f_0)$, i.e. a fixed point.

To show that the minimal fixed point is consistent, one works by induction on α. The only interesting part of this is to show that if $\Theta^\alpha(f_0)$ is consistent then so is $\Theta^{\alpha+1}(f_0)$. $\square$

---

[7]Remember "$\mathsf{P}_= (\ulcorner Blue \urcorner, \ulcorner 1/2 \urcorner)$" is an abbreviation for "$\mathsf{P}_\geqslant (\ulcorner Blue \urcorner, \ulcorner 1/2 \urcorner) \wedge \mathsf{P}_\geqslant (\ulcorner \neg Blue \urcorner, \ulcorner 1 - 1/2 \urcorner)$"

Definition 3.2.10. Let lfp denote the evaluation function which is the least fixed point of Θ.

If φ is grounded in facts that are not about truth or probability then this process of evaluation will terminate in such facts and the sentence will be evaluated appropriately in a fixed point. Such sentences will also therefore be given a point-valued probability as is desired. This will cover sentences that are expressible in the operator language, therefore showing that this semantics extends an operator semantics, a minimal adequacy requirement for any proposed semantics (see Section 3.3.1). However we will get more, for example $0 = 0 \vee \lambda$ isn't expressible in the operator language, but it will be evaluated positively in each world and so be assigned probability 1, i.e. $\mathsf{P}_= (\ulcorner 0 = 0 \vee \lambda \urcorner, \ulcorner 1 \urcorner)$ will also be evaluated positively.

The fixed points have some nice properties:

Proposition 3.2.11. *For f a fixed point of Θ we have:*

$$\#\varphi \in f(w) \iff (w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \varphi$$

Therefore we have

$$(w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{T}\ulcorner\varphi\urcorner \iff (w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \varphi$$
$$(w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\mathsf{T}\ulcorner\varphi\urcorner \iff (w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\varphi$$
$$(w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner) \iff m_w\{v \mid (v,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \varphi\} \geqslant r$$
$$(w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner) \iff m_w\{v \mid (v,f) \not\models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\varphi\} < r$$
$$\vdots$$

And other analogous properties as in Proposition 3.2.4.

Proof. Follows immediately from Definitions 3.2.3 and 3.2.5. $\qquad\square$

It is these fixed points which we propose as providing the (non-classical) semantics for the language. lfp is a particularly interesting one of these fixed points.

These will not be classical, i.e. there will be some φ, for example λ, such that $\neg\varphi \in f(w) \not\Longleftrightarrow \varphi \notin f(w)$ and $(w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\varphi \not\Longleftrightarrow (w,f) \not\models_{\mathfrak{M}}^{\mathsf{SKP}} \varphi$. If we do have some f which is classical for the sentences not involving truth[8] and a fixed point, we will exactly have a Prob-PW-model from Section 2.4. But we saw that for a frame that is omniscient, introspective, finite or countably additive, the frame will not support a Prob-PW-model, so there cannot be any such classical fixed point f in any of these frames.

3.2.3 The classical semantics

Particularly when we provide an axiomatisation, we will also be interested in a classical variant of these semantics. In this we use the interpretation of T and P that $(w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}}$ gives us to determine a classical model for the language $\mathcal{L}_{\mathsf{P}_{\geqslant},\mathsf{T}}$. This is common when working with Kripke's theory, the resulting model often

---

[8]In the sense that $(w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\varphi \iff (w,f) \not\models_{\mathfrak{M}}^{\mathsf{SKP}} \varphi$.

being called the "closed-off model". The axiomatisation that we will provide will be an axiomatisation in classical logic and will be axiomatising these classical variants of the semantics.

We will define *the induced model given by* $\mathfrak{M}$ *and* f *at* w, $\mathsf{IM}_{\mathfrak{M}}[w,f]$, by "closing off" the model by putting the unevaluated sentences outside of the extension of T and P. This is described pictorially by altering Fig. 3.2 to Fig. 3.3.

$\mathsf{IM}_{\mathfrak{M}}[w,f] \models \ldots$:
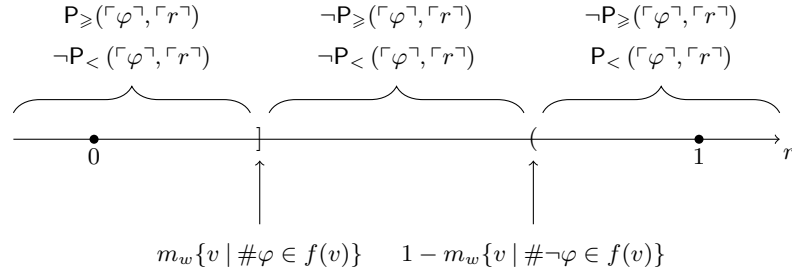
Figure 3.3: How $\mathsf{IM}_{\mathfrak{M}}[w,f]$ evaluates the probability of φ

It is defined formally as follows:

Definition 3.2.12. Define $\mathsf{IM}_{\mathfrak{M}}[w,f]$ to be a (classical) model for the language $\mathcal{L}_{\mathsf{P}_{\geqslant},\mathsf{T}}$ that has the domain $\mathbb{N}$, interprets the predicates from $\mathcal{L}$ as is specified by $\mathbf{M}(w)$, and interprets the other predicates by:

- $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}\overline{n} \iff (w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{T}\overline{n}$

- $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{P}_{\geqslant}(\overline{n},\overline{k}) \iff (w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{\geqslant}(\overline{n},\overline{k})$

This will satisfy:

Proposition 3.2.13. *For* $\mathfrak{M}$ *a probabilistic modal structure,* f *an evaluation function and* $w \in W$,

- $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{P}_{>}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner) \iff (w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{>}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)$

- $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{P}_{\leqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner) \iff (w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{\leqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)$

- $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{P}_{<}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner) \iff (w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{<}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)$

- $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{P}_{=}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner) \iff (w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{=}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)$

Although these equivalences hold, $\mathsf{IM}_{\mathfrak{M}}[w,f] \models$ differs from $(w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}}$ because $\mathsf{IM}_{\mathfrak{M}}[w,f]$ is classical, for example we might have that $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)$ but $(w,f) \not\models^{SKP}_{\mathfrak{M}} \neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)$.

As a corollary of the previous Proposition, we also have:

Corollary 3.2.14. *For* $\mathfrak{M}$ *a probabilistic modal structure,* f *a* fixed point *evaluation function and* $w \in W$,

- $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}\ulcorner\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)\urcorner \leftrightarrow \mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)$

- $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}\ulcorner\mathsf{P}_{>}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)\urcorner \leftrightarrow \mathsf{P}_{>}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)$

- *etc for all the positive principles.*

And for the negative principles, we have:

- $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}^\ulcorner \neg \mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)^\urcorner \leftrightarrow \mathsf{P}_{<}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)$

- *etc for all the negative principles, see Proposition 3.2.4.*

Proof. To show this, one starts by assuming the left hand side of each equivalence required, then uses Definition 3.2.12, then the result from Proposition 3.2.11 that for f a fixed point, $(w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{T}^\ulcorner\varphi\urcorner \iff (w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \varphi$ and then, at least for the negated principles, Proposition 3.2.4 and finally Proposition 3.2.13 to show that therefore the right hand side is satisfied in the induced model. $\qquad\square$

These induced models, for consistent fixed points f, are our proposal for the semantics of the language.

Although these models are models of classical logic, the truth and probability notions do not act classically. For example, for f a consistent fixed point we will have $\neg\mathsf{T}^\ulcorner\lambda\urcorner$, $\neg\mathsf{T}^\ulcorner\neg\lambda\urcorner$, $\neg\mathsf{P}_{>}(\ulcorner\lambda\urcorner, \ulcorner 0\urcorner)$ and $\neg\mathsf{P}_{<}(\ulcorner\lambda\urcorner, \ulcorner 1\urcorner)$ all satisfied in $\mathsf{IM}_{\mathfrak{M}}[w,f]$. Furthermore, some tautologies of classical logic are not in the extension of the truth predicate and not assigned probability 1, for example $\lambda \vee \neg\lambda$.[9]

3.2.4 P is an SK-probability

The underlying evaluation scheme we used to develop this construction was a strong Kleene scheme and we will show that the models we have developed can be seen to provide probabilities over logics arising from strong Kleene evaluations. So although traditional probabilism has been dropped, we have instead just modified it to be the appropriate version of probabilism over these non-classical logics.

One way in which the traditional probabilistic framework is rejected in our construction is that some sentences are assigned probability ranges. We therefore see that we have two functions to consider, $\underline{\mathrm{p}_{(w,f)}}$ and $\overline{\mathrm{p}_{(w,f)}}$:

Definition 3.2.15. Fix some probabilistic modal structure $\mathfrak{M}$, evaluation function f and world w. Define

$$\underline{\mathrm{p}_{(w,f)}}(\varphi) := \sup\{r \in \mathbb{Q} \mid \mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\}$$

$$\overline{\mathrm{p}_{(w,f)}}(\varphi) := \inf\{r \in \mathbb{Q} \mid \mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{P}_{\leqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\}$$

For f consistent we let p *as given by* $\mathfrak{M}$, w *and* f denote:

- $\underline{\mathrm{p}_{(w,f)}}(\varphi)$, if $\underline{\mathrm{p}_{(w,f)}}(\varphi) = \overline{\mathrm{p}_{(w,f)}}(\varphi)$,

- $[\underline{\mathrm{p}_{(w,f)}}(\varphi), \overline{\mathrm{p}_{(w,f)}}(\varphi)]$ otherwise.

This can be seen as in Fig. 3.4.

Using this definition (also by comparing Fig. 3.4 to Fig. 3.3) we have the following equivalent characterisations of $\overline{\mathrm{p}_{(w,f)}}$ and $\underline{\mathrm{p}_{(w,f)}}$

---

[9]In Chapter 4 we will consider a variant of the this construction where such classical-logical tautologies are assigned (point-valued) probability 1.

$\mathsf{IM}_{\mathfrak{M}}[w, f] \models$:
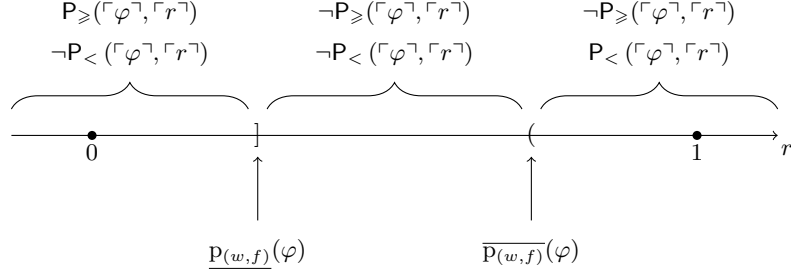
$$\underline{\mathrm{p}_{(w,f)}}(\varphi) \qquad \overline{\mathrm{p}_{(w,f)}}(\varphi)$$

Figure 3.4: Definition of $\underline{\mathrm{p}_{(w,f)}}(\varphi)$ and $\overline{\mathrm{p}_{(w,f)}}(\varphi)$

Proposition 3.2.16.

$$\underline{\mathrm{p}_{(w,f)}}(\varphi) = \sup\{r \mid (w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r \urcorner)\}$$
$$= m_w\{v \mid \#\varphi \in f(v)\}$$
$$\overline{\mathrm{p}_{(w,f)}}(\varphi) = \inf\{r \mid (w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_{\leqslant}(\ulcorner\varphi\urcorner, \ulcorner r \urcorner)\}$$
$$= 1 - m_w\{v \mid \#\neg\varphi \in f(v)\}$$

If f is a fixed point, then also,

$$\underline{\mathrm{p}_{(w,f)}}(\varphi) = m_w\{v \mid (w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \varphi\}$$
$$\overline{\mathrm{p}_{(w,f)}}(\varphi) = 1 - m_w\{v \mid (w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\varphi\}$$
$$= m_w\{v \mid (w, f) \not\models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\varphi\}$$

We will sometimes then say $\mathrm{p}(\varphi) = [\underline{\mathrm{p}_{(w,f)}}(\varphi), \overline{\mathrm{p}_{(w,f)}}(\varphi)]$, and if $\underline{\mathrm{p}_{(w,f)}}(\varphi) = \overline{\mathrm{p}_{(w,f)}}(\varphi) = r$ then we will say $\mathrm{p}(\varphi) = r$.

Both these functions lose nice properties one would expect from classical probabilities, for example $\underline{\mathrm{p}_{(w,f)}}(\lambda \vee \neg\lambda) = 0$, and $\overline{\mathrm{p}_{(w,f)}}(\lambda \wedge \neg\lambda) = 1$. However, one can show that $\underline{\mathrm{p}_{(w,f)}}$ is a non-classical probability over Kleene logic K_3, which is defined by truth preservation in Kleene evaluations, in the sense of Williams (2014), and $\overline{\mathrm{p}_{(w,f)}}$ is a non-classical probability over LP-logic, which is defined by falsity anti-preservation in Kleene evaluations. We will further analyse this connection in Section 7.3.4.

3.3 Connections to other languages

In this section we will present connections between this semantics and other languages. For general details about translations between such languages, see Section 1.6.3. There are two results we show in this section, the first is that the construction extends the operator semantics and the second is that a probability predicate can be reduced to a probability operator and a truth predicate.

3.3.1 Minimal adequacy of the theory

A minimal constraint on a predicate approach to probability is that it should be conservative over the operator approach. Our construction satisfies this constraint because if a sentence has a corresponding sentence in the operator language then the evaluation procedure for the sentence will terminate. Such sentences will therefore be assigned a truth value in the minimal fixed point, and so in any fixed point. Moreover it is easy to check that they receive the same truth value as the corresponding sentence in the operator approach.

For a language $\mathcal{L}$ as in Setup 3, consider the language $\mathcal{L}_{\mathbb{P}_{\geqslant r}, \mathbb{T}}$ defined as in Definition 1.6.17 And extend the definition of the semantics from Definition 2.2.1 by:

$$w \models_{\mathfrak{M}} \mathbb{T}\varphi \iff w \models_{\mathfrak{M}} \varphi$$

The language $\mathcal{L}_{\mathsf{P}_{\geqslant}, \mathsf{T}}$ 'extends' $\mathcal{L}_{\mathbb{P}_{\geqslant r}, \mathbb{T}}$ by the natural translation, ρ as defined in the next theorem.

Theorem 3.3.1. *There is some* $\rho : \mathsf{Sent}_{\mathcal{L}_{\mathbb{P}_{\geqslant r}, \mathbb{T}}} \to \mathsf{Sent}_{\mathcal{L}_{\mathsf{P}_{\geqslant}, \mathsf{T}}}$ *such that*

$$\rho(\varphi) = \begin{cases} \varphi & \varphi \in \mathcal{L} \\ \mathsf{P}_{\geqslant}(\ulcorner\rho(\varphi)\urcorner, \ulcorner r \urcorner) & \varphi = \mathbb{P}_{\geqslant r}(\psi) \\ \mathsf{T}\ulcorner\rho(\varphi)\urcorner & \varphi = \mathbb{T}\psi \\ \neg\rho(\psi) & \varphi = \neg\psi \\ \rho(\psi) \wedge \rho(\chi) & \varphi = \psi \wedge \chi \end{cases}$$

This theorem is a corollary of the result in Section 1.6.3.

Definition 3.3.2. For $\varphi \in \mathcal{L}_{\mathbb{P}_{\geqslant r}, \mathbb{T}}$, define $\mathsf{depth}(\varphi)$ to be:

- 0 if $\varphi \in \mathcal{L}$,

- $\mathsf{depth}(\psi) + 1$ for $\varphi = \mathbb{T}\psi$

- $\mathsf{depth}(\psi) + 1$ for $\varphi = \mathbb{P}_{\geqslant r}\psi$

- $\mathsf{depth}(\psi)$ for $\varphi = \neg\psi$

- $\max\{\mathsf{depth}(\psi), \mathsf{depth}(\chi)\}$ for $\varphi = \psi \vee \chi$

Theorem 3.3.3. *For each evaluation function* f,

$$w \models_{\mathfrak{M}} \varphi \iff (w, \Theta^{\mathsf{depth}(\varphi)}(f)) \models_{\mathfrak{M}}^{SKP} \rho(\varphi)$$

Proof. By induction on the positive complexity of φ.

φ in $\mathcal{L}$ is easy because then ρ is identity and the semantic clauses are the same.

Connective cases are fine.

$$\begin{aligned} w \models_{\mathfrak{M}} \mathbb{T}(\psi) \quad &\Longleftrightarrow \quad w \models_{\mathfrak{M}} \psi \\ &\Longleftrightarrow \quad (w, \Theta^{\mathsf{depth}(\psi)}(f)) \models_{\mathfrak{M}}^{SKP} \rho(\psi) \\ &\Longleftrightarrow \quad (w, \Theta^{\mathsf{depth}(\psi)+1}(f)) \models_{\mathfrak{M}}^{SKP} \mathsf{T}\ulcorner\rho(\psi)\urcorner \\ &\Longleftrightarrow \quad (w, \Theta^{\mathsf{depth}(\varphi)}(f)) \models_{\mathfrak{M}}^{SKP} \rho(\mathbb{T}(\psi)) \end{aligned}$$

$$
\begin{aligned}
w \models_{\mathfrak{M}} \neg \mathbb{T}(\psi) \quad &\Longleftrightarrow \quad w \models_{\mathfrak{M}} \neg \psi \\
&\Longleftrightarrow \quad (w, \Theta^{\mathsf{depth}(\psi)}(f)) \models_{\mathfrak{M}}^{SKP} \rho(\neg \psi) \\
&\Longleftrightarrow \quad (w, \Theta^{\mathsf{depth}(\psi)+1}(f)) \models_{\mathfrak{M}}^{SKP} \neg \mathsf{T} \ulcorner \rho(\psi) \urcorner \\
&\Longleftrightarrow \quad (w, \Theta^{\mathsf{depth}(\varphi)}(f)) \models_{\mathfrak{M}}^{SKP} \rho(\neg \mathbb{T}(\psi)) \\[6pt]
w \models_{\mathfrak{M}} \mathbb{P}_{\geqslant r}(\psi) \quad &\Longleftrightarrow \quad m_w\{v \mid v \models_{\mathfrak{M}} \psi\} \geqslant r \\
&\Longleftrightarrow \quad m_w\{v \mid (v, \Theta^{\mathsf{depth}(\psi)}(f)) \models_{\mathfrak{M}}^{SKP} \rho(\psi)\} \geqslant r \\
&\Longleftrightarrow \quad m_w\{v \mid \rho(\psi) \in \Theta^{\mathsf{depth}(\psi)+1}(f)(v)\} \geqslant r \\
&\Longleftrightarrow \quad (w, \Theta^{\mathsf{depth}(\varphi)}(f)) \models_{\mathfrak{M}}^{SKP} \mathsf{P}_{\geqslant}(\ulcorner \rho(\varphi) \urcorner, \ulcorner r \urcorner) \\[6pt]
w \models_{\mathfrak{M}} \neg \mathbb{P}_{\geqslant r}(\psi) \quad &\Longleftrightarrow \quad m_w\{v \mid v \models_{\mathfrak{M}} \psi\} < r \\
&\Longleftrightarrow \quad m_w\{v \mid v \models_{\mathfrak{M}} \neg \psi\} > 1 - r \\
&\Longleftrightarrow \quad m_w\{v \mid (v, \Theta^{\mathsf{depth}(\psi)}(f)) \models_{\mathfrak{M}}^{SKP} \rho(\neg \psi)\} > 1 - r \qquad \square \\
&\Longleftrightarrow \quad m_w\{v \mid \neg \rho(\psi) \in \Theta^{\mathsf{depth}(\psi)+1}(f)(v)\} > 1 - r \\
&\Longleftrightarrow \quad (w, \Theta^{\mathsf{depth}(\varphi)}(f)) \models_{\mathfrak{M}}^{SKP} \neg \mathsf{P}_{\geqslant}(\ulcorner \rho(\varphi) \urcorner, \ulcorner r \urcorner)
\end{aligned}
$$

For every $\varphi \in \mathsf{Sent}_{\mathcal{L}_{\mathbb{P}_{\geqslant r}, \mathbb{T}}}$, $\mathsf{depth}(\varphi) < \omega$, so:

Corollary 3.3.4. *For each evaluation function f,*

$$
w \models_{\mathfrak{M}} \varphi \iff (w, \Theta^{\omega}(f)) \models_{\mathfrak{M}}^{SKP} \rho(\varphi)
$$

Therefore, for every fixed point evaluation function f,

$$
w \models_{\mathfrak{M}} \varphi \iff (w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \rho(\varphi)
$$

3.3.2   Probability operators and a truth predicate

Halbach and Welch (2009) show that a modal operator and truth predicate are adequate for the language with the necessity predicate by translating a necessity predicate into "necessarily true", where this latter necessity is formalised as a modal operator, as standard in modal logic. Halbach and Welch view this as a defence of the operator approach against the charge of expressive weakness because this result shows that in fact the expressive power of a predicate language for necessity can be recovered within the operator setting if we also have a truth predicate. However, note that one has to still define a strong Kleene variant of the operator semantics and the resulting construction is just as complicated as the predicate construction. Stern (2015a) argues that this result can be alternatively viewed as a defence of the predicate approach against the backdrop of paradoxes, which in his context is particularly Montague's theorem, as it shows that the paradoxes can be *reduced* to paradoxes of truth.

We can show that a similar result holds in our construction.

Definition 3.3.5. Let $\mathcal{L}$ be any language as in Setup 3, let $\mathcal{L}_{\mathsf{T}}$ be the extension of that language with an added predicate, T. Let $\mathcal{L}_{\mathbb{P}_{\geqslant}, \mathsf{T}}$ be defined as in Definition 1.6.12 over the base language $\mathcal{L}_{\mathsf{T}}$, so adding operators $\mathbb{P}_{\geqslant}$ and allowing for quantifiers.

So there are sentences in this language like[10]

$$
\mathbb{P}_{\geqslant}(\mathsf{T} \ulcorner \forall x \mathbb{P}_{\geqslant}(\varphi(x), \ulcorner 1 \urcorner) \urcorner, \ulcorner 1/2 \urcorner).
$$

---

[10] As discussed in Section 1.6.3, the Gödel coding would code all sentences of the language $\mathcal{L}_{\mathbb{P}_{\geqslant}, \mathsf{T}}$ so allowing probabilities to appear inside the truth predicate.

We can construct a translation from the language with a probability predicate to one with a probability operator (and a truth predicate) by translating $\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)$ as $\mathbb{P}_{\geqslant}(\mathsf{T}\ulcorner\varphi\urcorner,\ulcorner r\urcorner)$ (at least for $\varphi \in \mathcal{L}$). This might be described as translating "the probability of φ is r" to "the probability that φ is true is r".

Theorem 3.3.6. *There is $\rho : \mathsf{Form}_{\mathsf{P}_{\geqslant},\mathsf{T}} \to \mathsf{Form}_{\mathbb{P}_{\geqslant},\mathsf{T}}$ such that*

$$
\rho(\varphi) = \begin{cases}
\varphi & \varphi \in \mathcal{L}_{\mathsf{T}} \ atomic \\
\mathbb{P}_{\geqslant}(\mathsf{T}\underset{.}{\rho}(t),s) & \varphi = \mathsf{P}_{\geqslant}(s,t) \\
\mathsf{T}\underset{.}{\rho}(t) & \varphi = \mathsf{T}t \\
\neg\rho(\psi) & \varphi = \neg\psi \\
\rho(\psi) \wedge \rho(\chi) & \varphi = \psi \wedge \chi \\
\forall x \rho(\psi) & \varphi = \forall x\psi \\
' & else
\end{cases}
$$

where $\underset{.}{\rho}$ is object level formula representing ρ, i.e. $\underset{.}{\rho}(\#\varphi) = \ulcorner\rho(\varphi)\urcorner$.

One can then construct a semantics for $\mathcal{L}_{\mathbb{P}_{\geqslant},\mathsf{T}}$ in the same way as we did for $\mathcal{L}_{\mathsf{P}_{\geqslant},\mathsf{T}}$, by altering $\models_{\mathfrak{M}}^{SKP}$ to $\models_{\mathfrak{M}}^{SK\mathbb{P}}$ as follows:

Definition 3.3.7. For $f \in \wp(\mathsf{cSent}_{\mathbb{P}_{\geqslant},\mathsf{T}})^{W}$ and $w \in W$, define $\models_{\mathfrak{M}}^{SK\mathbb{P}}$ inductively on the positive complexity of φ. Cases are as in the definition of $\models_{\mathfrak{M}}^{SKP}$ except the cases for $\mathsf{P}_{\geqslant}$ and $\neg\mathsf{P}_{\geqslant}$ are replaced by:

- $(w,f) \models_{\mathfrak{M}}^{SK\mathbb{P}} \mathbb{P}_{\geqslant}(\varphi,r)$ iff $m_w\{v \mid (v,f) \models_{\mathfrak{M}}^{SK\mathbb{P}} \varphi\} \geqslant r$

- $(w,f) \models_{\mathfrak{M}}^{SK\mathbb{P}} \neg\mathbb{P}_{\geqslant}(\varphi,r)$ iff $m_w\{v \mid (v,f) \models_{\mathfrak{M}}^{SK\mathbb{P}} \neg\varphi\} > 1 - r$

Definition 3.3.8. $\Theta_{\mathbb{P}}(f)(w) := \{\#\varphi \mid (w,f) \models_{\mathfrak{M}}^{SK\mathbb{P}} \varphi\}$

One can then observe that $\Theta_{\mathbb{P}}$ is monotone, and therefore that there is a fixed point of $\Theta_{\mathbb{P}}$.

We can then state the main result of this section.

Theorem 3.3.9. *If $\mathsf{lfp}_{\mathsf{P}}$ is the minimal fixed point of Θ and $\mathsf{lfp}_{\mathbb{P}}$ is the minimal fixed point of $\Theta_{\mathbb{P}}$ then*

$$(w,\mathsf{lfp}_{\mathsf{P}}) \models_{\mathfrak{M}}^{SK\mathsf{P}} \varphi \iff (w,\mathsf{lfp}_{\mathbb{P}}) \models_{\mathfrak{M}}^{SK\mathbb{P}} \rho(\varphi)$$

Proof. One can show that this in fact holds at all stages in the construction of lfp. Let $f_0^{\mathsf{P}}(w) = f_0^{\mathbb{P}} = \varnothing$ for all $w \in W$. Let f_α^{P} be the result of α-many applications of Θ to f_0^{P}, and similarly let $f_\alpha^{\mathbb{P}}$ denote α-many applications of $\Theta_{\mathbb{P}}$ to $f_0^{\mathbb{P}}$.

We will show that the equivalence holds at each α by induction on α. There must be some α where $f_\alpha^{\mathsf{P}} = \mathsf{lfp}_{\mathsf{P}}$ and $\Theta_\alpha^{\mathbb{P}} = \mathsf{lfp}_{\mathbb{P}}$ by standard results on inductive definitions (see, e.g., Moschovakis, 1974).

This holds for $\alpha = 0$ trivially.

For the successor step, the induction hypothesis is that

$$(w,f_\alpha^{\mathsf{P}}) \models_{\mathfrak{M}}^{SK\mathsf{P}} \varphi \iff (w,f_\alpha^{\mathbb{P}}) \models_{\mathfrak{M}}^{SK\mathbb{P}} \rho(\varphi)$$

and we need to show this equivalence for $\alpha + 1$ which we do by a sub-induction on the positive complexity of φ.

For t a closed term we have: If there is no $\varphi \in \mathsf{Sent}_{\mathsf{P}_{\geqslant},\mathsf{T}}$ such that $t^{\mathbb{N}} = \#\varphi$ then:

$$(w, f^{\mathsf{P}}_{\alpha+1}) \models^{SK\mathsf{P}}_{\mathfrak{M}} \neg \mathsf{T}t,$$
$$\text{and } (w, f^{\mathbb{P}}_{\alpha+1}) \models^{SK\mathbb{P}}_{\mathfrak{M}} \neg \rho(\mathsf{T}t),$$
$$\text{and } (w, f^{\mathsf{P}}_{\alpha+1}) \models^{SK\mathsf{P}}_{\mathfrak{M}} \neg \mathsf{P}_{\geqslant}(t, s),$$
$$\text{and } (w, f^{\mathbb{P}}_{\alpha+1}) \models^{SK\mathbb{P}}_{\mathfrak{M}} \neg \rho(\mathsf{P}_{\geqslant}(t, s)).$$

For s a closed term we have: If there is no r such that $s^{\mathbb{N}} = \#r$ then:

$$(w, f^{\mathsf{P}}_{\alpha+1}) \models^{SK\mathsf{P}}_{\mathfrak{M}} \neg \mathsf{P}_{\geqslant}(t, s),$$
$$\text{and } (w, f^{\mathbb{P}}_{\alpha+1}) \models^{SK\mathbb{P}}_{\mathfrak{M}} \neg \rho(\mathsf{P}_{\geqslant}(t, s)).$$

So we only need to consider the cases for $t^{\mathbb{N}} = \#\varphi$ and $s^{\mathbb{N}} = \#r$. All the cases for $\mathsf{T}\ulcorner\varphi\urcorner$, $\neg\mathsf{T}\ulcorner\varphi\urcorner$, $\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)$ and $\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)$ are similar. We just present the cases for $\mathsf{T}\ulcorner\varphi\urcorner$ and $\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)$

$$(w, f^{\mathsf{P}}_{\alpha+1}) \models^{SK\mathsf{P}}_{\mathfrak{M}} \mathsf{T}\ulcorner\varphi\urcorner$$
$$\iff \#\varphi \in f^{\mathsf{P}}_{\alpha+1}(w)$$
$$\iff (w, f^{\mathsf{P}}_{\alpha}) \models^{SK\mathsf{P}}_{\mathfrak{M}} \varphi$$
$$\iff (w, f^{\mathbb{P}}_{\alpha}) \models^{SK\mathbb{P}}_{\mathfrak{M}} \rho(\varphi)$$
$$\iff \#\rho(\varphi) \in f^{\mathbb{P}}_{\alpha+1}(w)$$
$$\iff (w, f^{\mathbb{P}}_{\alpha+1}) \models^{SK\mathbb{P}}_{\mathfrak{M}} \mathsf{T}\ulcorner\rho(\varphi)\urcorner$$
$$\iff (w, f^{\mathbb{P}}_{\alpha+1}) \models^{SK\mathbb{P}}_{\mathfrak{M}} \rho(\mathsf{T}\ulcorner\varphi\urcorner)$$

$$(w, f^{\mathsf{P}}_{\alpha+1}) \models^{SK\mathsf{P}}_{\mathfrak{M}} \neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)$$
$$\iff m_w\{v \mid \#\neg\varphi \in f^{\mathsf{P}}_{\alpha+1}(v)\} > 1 - r$$
$$\iff m_w\{v \mid (v, f^{\mathsf{P}}_{\alpha}) \models^{SK\mathsf{P}}_{\mathfrak{M}} \neg\varphi\} > 1 - r$$
$$\iff m_w\{v \mid (v, f^{\mathbb{P}}_{\alpha}) \models^{SK\mathbb{P}}_{\mathfrak{M}} \neg\rho(\varphi)\} > 1 - r$$
$$\iff m_w\{v \mid (v, f^{\mathbb{P}}_{\alpha+1}) \models^{SK\mathbb{P}}_{\mathfrak{M}} \neg\mathsf{T}\ulcorner\rho(\varphi)\urcorner\} > 1 - r$$
$$\iff (w, f^{\mathbb{P}}_{\alpha+1}) \models^{SK\mathbb{P}}_{\mathfrak{M}} \neg\mathbb{P}_{\geqslant}(\mathsf{T}\ulcorner\rho(\varphi)\urcorner, \ulcorner r\urcorner)$$
$$\iff (w, f^{\mathbb{P}}_{\alpha+1}) \models^{SK\mathbb{P}}_{\mathfrak{M}} \rho(\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner))$$

The inductive steps for this sub-induction on the positive complexity of φ are simple because the definition of $\models^{SK\mathsf{P}}_{\mathfrak{M}}$ and $\models^{SK\mathbb{P}}_{\mathfrak{M}}$ are exactly the same for these and the translation function commutes with these operations.

Now suppose μ is a limit ordinal. Then the argument is very similar to the successor stage and again works by a subinduction on the positive complexity

of φ. For example for the base case $\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)$:

$$(w, f_\mu^{\mathsf{P}}) \models_{\mathfrak{M}}^{SK\mathsf{P}} \mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)$$

$$\Longleftrightarrow m_w\{v \mid \#\varphi \in f_\mu^{\mathsf{P}}(v)\} \geqslant r$$

$$\Longleftrightarrow m_w\left(\bigcup_{\alpha<\mu}\{v \mid \#\varphi \in f_\alpha^{\mathsf{P}}(v)\}\right) \geqslant r$$

$$\Longleftrightarrow m_w\left(\bigcup_{\alpha<\mu}\{v \mid (v, f_\alpha^{\mathsf{P}}) \models_{\mathfrak{M}}^{SK\mathsf{P}} \mathsf{T}\ulcorner\varphi\urcorner\}\right) \geqslant r$$

$$\Longleftrightarrow m_w\left(\bigcup_{\alpha<\mu}\{v \mid (v, f_\alpha^{\mathbb{P}}) \models_{\mathfrak{M}}^{SK\mathbb{P}} \mathsf{T}\ulcorner\rho(\varphi)\urcorner\}\right) \geqslant r$$

$$\Longleftrightarrow m_w\left(\bigcup_{\alpha<\mu}\{v \mid \#\rho(\varphi) \in \Theta_{\mathbb{P}}^\alpha(f_0)\}\right) \geqslant r$$

$$\Longleftrightarrow m_w\{v \mid \#\rho(\varphi) \in \Theta_{\mathbb{P}}^\mu(f_0)\} \geqslant r$$

$$\Longleftrightarrow m_w\{v \mid (v, f_\mu^{\mathbb{P}}) \models_{\mathfrak{M}}^{SK\mathbb{P}} \mathsf{T}\ulcorner\rho(\varphi)\urcorner\} \geqslant r$$

$$\Longleftrightarrow (w, f_\mu^{\mathbb{P}}) \models_{\mathfrak{M}}^{SK\mathbb{P}} \mathbb{P}_{\geqslant}(\mathsf{T}\ulcorner\rho(\varphi)\urcorner, \ulcorner r\urcorner)$$

$$\Longleftrightarrow (w, f_\mu^{\mathbb{P}}) \models_{\mathfrak{M}}^{SK\mathbb{P}} \rho(\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)) \qquad\qquad \square$$

3.4 Specific cases of the semantics

We will now discuss what happens when we consider particular probabilistic modal structures. In doing this we consider the property of introspection and $\mathbb{N}$-additivity, the latter holding in probabilistic modal structures that have countably additive accessibility measures.

3.4.1 Introspection

Studying introspection in languages that allow for self-referential probabilities is interesting because if it is naively formulated it is inconsistent, a problem discussed by Caie (2013) and Christiano et al. (ms) and presented in Section 1.4.

Introspective probabilistic modal structures satisfy

$$\mathbb{P}_{\geqslant r}\varphi \to \mathbb{P}_{=1}\mathbb{P}_{\geqslant r}\varphi$$
$$\text{and } \neg\mathbb{P}_{\geqslant r}\varphi \to \mathbb{P}_{=1}\neg\mathbb{P}_{\geqslant r}\varphi$$

Introspective structures do not satisfy the direct translations of these in our semantics:

$$\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner) \to \mathsf{P}_{=}(\ulcorner\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner, \ulcorner 1\urcorner)$$
$$\text{and } \neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner) \to \mathsf{P}_{=}(\ulcorner\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner, \ulcorner 1\urcorner)$$

But they *do* satisfy varients of these expressed using a truth predicate:

$$\mathsf{T}\ulcorner\mathsf{P}_{\geqslant}(\varphi, \ulcorner r\urcorner)\urcorner \implies \mathsf{P}_{=}(\ulcorner\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner, \ulcorner 1\urcorner)$$
$$\text{and } \mathsf{T}\ulcorner\neg\mathsf{P}_{\geqslant}(\varphi, \ulcorner r\urcorner)\urcorner \implies \mathsf{P}_{=}(\ulcorner\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner, \ulcorner 1\urcorner)$$

Proposition 3.4.1. *Let $\mathfrak{M}$ be weakly introspective (a mild weakening of the condition: $m_w\{v \mid m_v = m_w\} = 1$ for all w). Then for any evaluation function f and world w,*

- *If $(w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)$,*
  then $(w,\Theta(f)) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_{=}(\ulcorner\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner, \ulcorner 1\urcorner)$

- *If $(w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)$,*
  then $(w,\Theta(f)) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_{=}(\ulcorner\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner, \ulcorner 1\urcorner)$

And similarly for $\mathsf{P}_{>}, \mathsf{P}_{\leqslant}$ etc.
  By the definition of $\mathsf{IM}_{\mathfrak{M}}[w,f]$ we therefore have: [11]

- $\mathsf{IM}_{\mathfrak{M}}[w,\Theta(f)] \models \mathsf{T}\ulcorner\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner \to \mathsf{P}_{=}(\ulcorner\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner, \ulcorner 1\urcorner))$

- $\mathsf{IM}_{\mathfrak{M}}[w,\Theta(f)] \models \mathsf{T}\ulcorner\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner \to \mathsf{P}_{=}(\ulcorner\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner, \ulcorner 1\urcorner))$

And similarly for $\mathsf{P}_{>}$ etc.
  Therefore for f a fixed point evaluation function,

- $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}\ulcorner\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner \to \mathsf{P}_{=}(\ulcorner\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner, \ulcorner 1\urcorner))$

- $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}\ulcorner\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner \to \mathsf{P}_{=}(\ulcorner\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner, \ulcorner 1\urcorner))$

And similarly for $\mathsf{P}_{>}$ etc.

Proof.

$$
\begin{aligned}
(w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner) &\iff m_w\{v \mid \#\varphi \in f(v)\} \geqslant r \\
&\implies m_w\{v' \mid m_{v'}\{v \mid \#\varphi \in f(v)\} \geqslant r\} = 1 \\
&\iff m_w\{v' \mid (v',f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\} = 1 \\
&\iff m_w\{v' \mid \#\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner) \in \Theta(f)\} = 1 \\
&\iff (w,\Theta(f)) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_{=}(\ulcorner\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner, \ulcorner 1\urcorner)
\end{aligned}
$$

$$
\begin{aligned}
(w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner) &\iff m_w\{v \mid \#\neg\varphi \in f(v)\} > 1 - r \\
&\implies m_w\{v' \mid m_{v'}\{v \mid \#\neg\varphi \in f(v)\} > 1 - r\} = 1 \\
&\iff m_w\{v' \mid (v',f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\} = 1 \\
&\iff m_w\{v' \mid \#\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner) \in \Theta(f)\} = 1 \\
&\iff (w,\Theta(f)) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_{=}(\ulcorner\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner, \ulcorner 1\urcorner)
\end{aligned}
$$

And similar arguments hold for $\mathsf{P}_{>}, \mathsf{P}_{\leqslant}$ etc. The other characterisations are just easy consequences of this. $\qquad\square$

We also have the converse result which works very similarly to Proposition 2.3.2.

---

[11] In fact the quantified versions

- $\mathsf{IM}_{\mathfrak{M}}[w,\Theta(\Theta(f))] \models \forall a \forall x (\mathsf{T}\mathsf{P}_{\dot{\geqslant}}(x,a) \to \mathsf{P}_{\geqslant}(\mathsf{P}_{\dot{\geqslant}}(x,a), \ulcorner 1\urcorner))$

- $\mathsf{IM}_{\mathfrak{M}}[w,\Theta(\Theta(f))] \models \forall a \forall x (\mathsf{T}\neg\mathsf{P}_{\dot{\geqslant}}(x,a) \to \mathsf{P}_{\geqslant}(\neg\mathsf{P}_{\dot{\geqslant}}(x,a), \ulcorner 1\urcorner))$

are satisfied but we do not present this because it is not important for our point.

3.4 Specific cases of the semantics

Proposition 3.4.2. *Let $\mathcal{L}$ contain at least one empirical symbol. Then a probabilistic modal frame $(W, \{m_w\})$ is weakly introspective if and only if for every $\mathfrak{M}$ based on it, and every fixed point evaluation function f,*

- $\mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{T}\ulcorner\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner \to \mathsf{P}_{=}(\ulcorner\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner, \ulcorner 1\urcorner))$

- $\mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{T}\ulcorner\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner \to \mathsf{P}_{=}(\ulcorner\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner, \ulcorner 1\urcorner))$

or, equivalently,

- *If $(w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)$,*
  then $(w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{=}(\ulcorner\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner, \ulcorner 1\urcorner)$

- *If $(w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)$,*
  then $(w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{=}(\ulcorner\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner, \ulcorner 1\urcorner)$

Proof. We have shown $\implies$ in Proposition 3.4.1. The "equivalently" works by Propositions 3.2.11 and 3.2.13.

$\impliedby$: Suppose the RHS. Fix w and some $A \subseteq W$. WLOG suppose $\mathcal{L}$ contains the propositional variable O. Consider a valuation $\mathbf{M}$ such that

$$\mathbf{M}(v) \models O \iff v \in A.$$

Observe that for f a fixed point,[12]

$$\#O \in f(v) \iff v \in A$$
$$\#\neg O \in f(v) \iff v \notin A.$$

Suppose $r \leqslant m_w(A) < q$.

Then $(w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{\geqslant}(\ulcorner O\urcorner, \ulcorner r\urcorner)$ and $(w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\mathsf{P}_{\geqslant}(\ulcorner O\urcorner, \ulcorner q\urcorner)$. So

$$(w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{\geqslant}(\ulcorner\mathsf{P}_{\geqslant}(\ulcorner O\urcorner, \ulcorner r\urcorner)\urcorner, \ulcorner 1\urcorner)$$

so $\quad m_w\{v \mid \#\mathsf{P}_{\geqslant}(\ulcorner O\urcorner, \ulcorner r\urcorner) \in f(v)\} \geqslant 1$

so $\quad m_w\{v \mid (v, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{\geqslant}(\ulcorner O\urcorner, \ulcorner r\urcorner)\} \geqslant 1$

so $\quad m_w\{v \mid m_v\{v' \mid \#O \in f(v')\} \geqslant r\} \geqslant 1$

so $\quad m_w\{v \mid m_v(A) \geqslant r\} \geqslant 1$

and

$$(w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{\geqslant}(\ulcorner\neg\mathsf{P}_{\geqslant}(\ulcorner O\urcorner, \ulcorner q\urcorner)\urcorner, \ulcorner 1\urcorner)$$

so $\quad m_w\{v \mid \#\neg\mathsf{P}_{\geqslant}(\ulcorner O\urcorner, \ulcorner q\urcorner) \in f(v)\} \geqslant 1$

so $\quad m_w\{v \mid (v, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\mathsf{P}_{\geqslant}(\ulcorner O\urcorner, \ulcorner q\urcorner)\} \geqslant 1$

so $\quad m_w\{v \mid m_v\{v' \mid \#\neg O \in f(v')\} > 1 - q\} \geqslant 1$

so $\quad m_w\{v \mid m_v(W \setminus A) > 1 - q\} \geqslant 1$

so $\quad m_w\{v \mid m_v(A) < q\} \geqslant 1$

Since m_w is finitely additive, we therefore have:

$$m_w\{v \mid r \leqslant m_v(A) < q\} = 1 \qquad\qquad \square$$

---

[12]In fact just requires that $f = \Theta(g)$ for some g.

75

In fact in this setting, we can also formulate a version of these introspection principles that does not involve reference to the truth predicate. That is by the principles:

$$\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner) \to \mathsf{P}_{=}(\ulcorner\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)\urcorner,\ulcorner 1\urcorner)$$
$$\mathsf{P}_{<}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner) \to \mathsf{P}_{=}(\ulcorner\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)\urcorner,\ulcorner 1\urcorner) \tag{3.1}$$

Which are equivalent to the principles:[13]

$$\mathsf{T}\ulcorner\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)\urcorner \to \mathsf{P}_{=}(\ulcorner\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)\urcorner,\ulcorner 1\urcorner)$$
$$\mathsf{T}\ulcorner\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)\urcorner \to \mathsf{P}_{=}(\ulcorner\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)\urcorner,\ulcorner 1\urcorner)$$

The reason that Eq. (3.1) is weaker than the principle which cannot be satisfied:

$$\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner) \to \mathsf{P}_{=}(\ulcorner\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)\urcorner,\ulcorner 1\urcorner) \tag{3.2}$$

is that for the problematic sentences $(w,f) \nvDash_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)$ and $(w,f) \nvDash_{\mathfrak{M}}^{\mathsf{SKP}}$ $\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)$, so neither antecedent of the consistent principles will be satisfied in $\mathsf{IM}_{\mathfrak{M}}[w,f]$, though $\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)$ will be.

This bears some connection to discussions in epistemology where the negative introspection principles are rejected, often because of the possibility of suspending judgement. Stating our antecedents with the truth predicate means that the introspection principle will not be applied to sentences where neither $\mathsf{T}\ulcorner\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)\urcorner$ nor $\mathsf{T}\ulcorner\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner,\ulcorner r\urcorner)\urcorner$. So whereas the usual rejection of introspection involves a suspension of judgement by the agent, our alteration here is due to the "suspension" by the truth predicate.

Although we have this alternative form without the truth predicate in this setting we will not in general focus on that as a solution because it is not systematic and general. As we will see in Chapter 5, and in particular Section 5.4.2, the strategy of expressing the principles using a truth predicate is a general strategy and also applies when other theories of truth are being considered as.

This strategy can also be applied to other principles, as we will discuss in Section 5.4.2, though further work should be done to give a general result about how to express these principles.

This strategy comes from Stern (2014a) where he applies the strategy in frameworks where all-or-nothing modalities are conceived of as predicates. In that article he describes the strategy as "avoiding introduction and elimination of the modal predicate independently of the truth predicate" and suggests that this might in general allow one to avoid inconsistency and paradox, at least if one accepts a consistent theory of truth. Moreover he says:

> [This strategy] seems to be well motivated if one adopts the deflationist idea that quotation and disquotation are the function of the truth predicate. Consequently, quotation and disquotation of sentences is not the task of the modal predicate and in formulating modal principles we should therefore avoid the introduction or elimination of the modal predicates without the detour via the truth predicate. (Stern, 2014a, p. 227)

---

[13]This can be seen using Corollary 3.2.14.

That, then, is our proposal for expressing the introspection principles. The result that our formulation of the introspection principles exactly pin down the weakly introspective probabilistic modal structures is a very strong argument in favour of this formulation of the principles.

We next show a nice feature of the construction, namely that it can account for ℕ-additivity. This will not follow the strategy outlined in the preceeding section because in ℕ-additivity there is no introduction or elimination of the modal notions. However, there are still some points to note because we can interpret our semantics as assigning sentences probability *ranges* instead of point values.

3.4.2 ℕ-additivity

We introduced ℕ-additivity in Section 1.2.1, it said

$$\mathrm{p}(\exists x \varphi(x)) = \lim_n \mathrm{p}(\varphi(\overline{0}) \vee \ldots \vee \varphi(\overline{n}))$$

and captures the idea that p is countably additive and that the domain is ℕ. In our semantics here, sentences are sometimes given ranges of probability values instead of points, so the definition of ℕ-additivity needs to be reformulated to account for this.

Previous work on self-referential probability has faced a challenge from ℕ-additivity, both Christiano's requirements (Christiano et al., ms) and the final stage of Leitgeb's construction (Leitgeb, 2012) lead to a formula $\varphi(x)$ such that for each n $\mathrm{p}(\varphi(\overline{0}) \vee \ldots \vee \varphi(\overline{n})) = 0$ but $\mathrm{p}(\exists x \varphi(x)) = 1$. [14] A big difference between our semantics in this chapter and their analysis is that they always assumed that the probabilities assigned point values to each sentence (satisfying the usual classical probability axioms), whereas we allow for these ranges. In this way we are able to avoid the problems that they faced.

In fact, as a consequence of our result in Section 2.4.2, we see that this use of ranges is essential in frames where each $\mathbf{M}(w)$ is an ℕ-model (which is an assumption for all our frames in this chapter) and each m_w is countably additive. In that section we showed that there were no Prob-PW-models for such structures, so instead there must be some sentences which do not receive a single point valued probability. This is because we are provided with a Prob-PW-model by finding a fixed point evaluation function where for every sentence, φ there is some r where $\mathsf{P}_=(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)$ is satisfied.

The way we reformulate the ℕ-additivity criterion is to apply it to the end points of the ranges of probability values assigned.

Definition 3.4.3. We say that p as given by $\mathfrak{M}, w, f$ is ℕ-*additive* if

$$\underline{\mathrm{p}_{(w,f)}}(\exists x \varphi(x)) = \lim_n \underline{\mathrm{p}_{(w,f)}}(\varphi(\overline{0}) \vee \ldots \vee \varphi(\overline{n}))$$

---

[14]The failure of ℕ-additivity in Leitgeb's theory is closely related to McGee's ω-inconsistency result from McGee (1985), presented in Theorem 2.4.12. For Leitgeb a sentence displaying the bad failure of ℕ-additivity is: $\neg \overbrace{\mathsf{T}\ulcorner\mathsf{T}\ulcorner\ldots\mathsf{T}\ulcorner\delta\urcorner\ldots\urcorner\urcorner}^{n+1}$ where δ is the McGee sentence with truth, namely is a sentence with the property that $\mathsf{PA} \vdash \delta \leftrightarrow \exists n \neg \overbrace{\mathsf{T}\ulcorner\mathsf{T}\ulcorner\ldots\mathsf{T}\ulcorner\delta\urcorner\ldots\urcorner\urcorner}^{n+1}$. For Christiano et al. this is given by $\mathsf{P}\ulcorner\epsilon\urcorner \leqslant 1 - 1/n+1$ where ϵ is a sentence with the property $\mathsf{PA} \vdash \epsilon \leftrightarrow \mathsf{P}\ulcorner\epsilon\urcorner < 1$; the fact that Christiano et al. face a challenge from ℕ-additivity was pointed out to me by Hannes Leitgeb.

and similarly for $\overline{\mathrm{p}_{(w,f)}}$.

If we consider a probabilistic modal structure where the measure m_w is countably additive then p will be $\mathbb{N}$-additive.

Theorem 3.4.4. *If $\mathfrak{M}$ is such that m_w is countably additive,[15] and f is a fixed point, then P as given by $\mathfrak{M}, w, f$ will be $\mathbb{N}$-additive.*

Proof. By the definition of $(w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}}$, we have:

$$\{v \mid (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \exists x \varphi(x)\} = \bigcup_n \{v \mid (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \varphi(\overline{0}) \vee \ldots \vee \varphi(\overline{n})\}$$

So since m_w is countably additive,

$$m_w\{v \mid (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \exists x \varphi(x)\} = \lim_n m_w\{v \mid (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \varphi(\overline{0}) \vee \ldots \vee \varphi(\overline{n})\}$$

This suffices for the result for $\underline{\mathrm{p}_{(w,f)}}$ because by Proposition 3.2.16 we have

$$\underline{\mathrm{p}_{(w,f)}}(\psi) = m_w\{v \mid (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \psi\}.$$

$$\{v \mid (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\exists x \varphi(x)\} = \bigcap_n \{v \mid (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg(\varphi(\overline{0}) \vee \ldots \vee \varphi(\overline{n}))\}$$

So

$$m_w\{v \mid (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\exists x \varphi(x)\} = \lim_n m_w\{v \mid (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg(\varphi(\overline{0}) \vee \ldots \vee \varphi(\overline{n}))\}$$

This suffices for the result for $\underline{\mathrm{p}_{(w,f)}}$ because by Proposition 3.2.16 we have

$$\overline{\mathrm{p}_{(w,f)}}(\psi) = 1 - m_w\{v \mid (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\psi\}. \qquad \square$$

This is the form of $\mathbb{N}$-additivity that is appropriate in a context where one deals with interval-valued probabilities. We therefore see that if we don't restrict ourselves to merely finitely-additive probabilities then we can account for $\mathbb{N}$-additivity.

3.4.3 This extends the usual truth construction

This construction we have presented extends the usual construction just for the truth predicate. In fact, if **M** is constant, i.e. assigns the same valuation to each world, then the construction is just equivalent to the construction for truth. In that case sentences either receive probability 1, if they are true in the usual Kripkean construction, or 0 if they are false, or $[0, 1]$ if neither.

Theorem 3.4.5. [16] *If $\mathfrak{M}$ is a probabilistic modal structure with **M** constant,[17] then for each w*

---

[15] We can drop the condition that m_w be defined on the whole powerset of W and instead ask just that it is defined an algebra of subsets containing the sets of the form $\{v \mid n \in f(v)\}$.

[16] Question was from Stanislav Speranski.

[17] I.e. for all $w, v \in W$, $\mathbf{M}(w) = \mathbf{M}(v)$.

1. $(w, \mathsf{lfp}) \models^{\mathsf{SKP}}_{\mathfrak{M}} \varphi \iff (v, \mathsf{lfp}) \models^{SKP}_{\mathfrak{M}} \varphi$, *for all* v,

2. *Let* p *be as given by* $\mathfrak{M}$, w *and* lfp *(see Definition 3.2.15). For all* $\varphi \in$ $\mathsf{Sent}_{\mathsf{P}_{\geqslant},\mathsf{T}}$, *either* $\mathrm{p}(\varphi) = 0$, $\mathrm{p}(\varphi) = 1$ *or* $\mathrm{p}(\varphi) = [0, 1]$,

3. $\mathrm{p}(\varphi) = 1 \iff (w, \mathsf{lfp}) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{T}\ulcorner\varphi\urcorner$

4. $\mathsf{IM}_{\mathfrak{M}}[w, \mathsf{lfp}] \models \forall x (\mathsf{P}_{>}(x, \ulcorner 0 \urcorner) \leftrightarrow \mathsf{T}x)$

Proof. We can prove that these the first hold for each f_α in the construction of the least fixed point. For the induction step we do a sub-induction on the positive complexity of φ.

The other properties also hold for each stage and follow from the first property. $\square$

Definition 3.4.6. Define $\rho : \mathsf{Sent}_{\mathsf{P}_{\geqslant},\mathsf{T}} \to \mathsf{Sent}_{\mathsf{T}}$ with

$$
\rho(\varphi) = \begin{cases}
\varphi & \varphi \in \mathcal{L} \text{ atomic} \\
\mathsf{T}\dot{\rho}(t) & \varphi = \mathsf{T}t \\
s \preceq \ulcorner 0 \urcorner \vee (s \succ \ulcorner 0 \urcorner \wedge \mathsf{T}\dot{\rho}(t)) & \varphi = \mathsf{P}_{\geqslant}(t, s) \\
\neg \rho(\psi) & \varphi = \neg\psi \\
\rho(\psi) \wedge \rho(\chi) & \varphi = \psi \wedge \chi \\
\forall x \rho(\psi) & \varphi = \forall x\psi \\
' & \text{else}
\end{cases}
$$

Such a ρ exists.

Theorem 3.4.7. *Then* $\#\varphi \in \mathsf{lfp}(w) \iff \#\rho(\varphi) \in \mathsf{lfp}_{\mathsf{T}}$.

Proof. We induct on the construction of the fixed point as in Theorem 3.3.9. As in Theorem 3.3.9, let $f^{\mathsf{P}}_0(w) = \varnothing = f^{\mathsf{T}}_0$ and let f^{P}_α denote the evaluation function from α-many applications of Θ to f^{P}_0. Let f^{T}_α denote the subset of $\mathbb{N}$ resulting from α-many applications of Γ to $\varnothing$.

We will show: $n \in f^{\mathsf{P}}_\alpha(w) \iff \dot{\rho}n^{\mathbb{N}} \in f^{\mathsf{T}}_\alpha$ for all r and w. For the induction step: Suppose it holds for r, then it suffices to show that

$$(w, f^{\mathsf{P}}_\alpha) \models^{\mathsf{SKP}}_{\mathfrak{M}} \iff f^{\mathsf{T}}_\alpha \models^{\mathsf{SKT}}_{\mathfrak{M}} \varphi$$

by induction on the positive complexity of φ. If $\varphi \in \mathcal{L}$ is atomic or negated atomic then

$$(w, f^{\mathsf{P}}_\alpha) \models^{\mathsf{SKP}}_{\mathfrak{M}} \varphi \iff \mathbf{M}(w) \models \varphi \iff \mathbf{M}_{\mathsf{T}} \models \varphi \iff f^{\mathsf{P}}_\alpha \models^{\mathsf{SKT}}_{\mathfrak{M}} \varphi$$

For $\varphi = \mathsf{T}t$,

$$
\begin{aligned}
(w, f^{\mathsf{P}}_\alpha) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{T}t \quad &\iff t^{\mathbb{N}} \in f^{\mathsf{P}}_\alpha(w) \text{ and } t^{\mathbb{N}} \in \mathsf{Sent}_{\mathcal{L}_{\mathsf{P}_{\geqslant},\mathsf{T}}} \\
&\iff \dot{\rho}t^{\mathbb{N}} \in f^{\mathsf{T}}_\alpha \text{ and } \dot{\rho}t^{\mathbb{N}} \in \mathsf{Sent}_{\mathcal{L}_{\mathsf{T}}} \\
&\iff f^{\mathsf{P}}_\alpha \models^{\mathsf{SKT}}_{\mathfrak{M}} \mathsf{T}\dot{\rho}t
\end{aligned}
$$

Similarly for $\neg\mathsf{T}t$

For $\varphi = \mathsf{P}_{\geqslant}(t,s)$, suppose $s^{\mathbb{N}} \succ 0$.

$$
\begin{aligned}
(w, f_\alpha^{\mathsf{P}}) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_{\geqslant}(t,s) \quad &\Longleftrightarrow \quad s^{\mathbb{N}} \in \mathsf{Rat} \text{ and } m_w\{v \mid t^{\mathbb{N}} \in f_\alpha^{\mathsf{P}}(v)\} \geqslant \mathsf{rat}(s^{\mathbb{N}}) \\
&\Longleftrightarrow \quad t^{\mathbb{N}} \in f_\alpha^{\mathsf{P}}(w) \\
&\Longleftrightarrow \quad \rho t^{\mathbb{N}} \in f_\alpha^{\mathsf{T}} \\
&\Longleftrightarrow \quad f_\alpha^{\mathsf{P}} \models_{\mathfrak{M}}^{\mathsf{SKT}} \mathsf{T}\rho t \\
&\Longleftrightarrow \quad f_\alpha^{\mathsf{P}} \models_{\mathfrak{M}}^{\mathsf{SKT}} s \preceq \ulcorner 0 \urcorner \vee (s \succ \ulcorner 0 \urcorner \wedge \mathsf{T}\dot{\rho}t)
\end{aligned}
$$

Suppose $s^{\mathbb{N}} \preceq 0$.

$$
\begin{aligned}
(w, f_\alpha^{\mathsf{P}}) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_{\geqslant}(t,s) \quad &\Longleftrightarrow \quad s^{\mathbb{N}} \in \mathsf{Rat} \text{ and } m_w\{v \mid t^{\mathbb{N}} \in f_\alpha^{\mathsf{P}}(v)\} \geqslant \mathsf{rat}(s^{\mathbb{N}}) \\
&\Longleftrightarrow \quad \top \\
&\Longleftrightarrow \quad f_\alpha^{\mathsf{P}} \models_{\mathfrak{M}}^{\mathsf{SKT}} s \preceq \ulcorner 0 \urcorner \vee (s \succ \ulcorner 0 \urcorner \wedge \mathsf{T}\dot{\rho}t)
\end{aligned}
$$

For $s^{\mathbb{N}} \notin \mathsf{Rat}$, $(w, f_\alpha^{\mathsf{P}}) \not\models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_{\geqslant}(t,s)$ and $f_\alpha^{\mathsf{P}} \not\models_{\mathfrak{M}}^{\mathsf{SKT}} s \preceq \ulcorner 0 \urcorner \vee (s \succ \ulcorner 0 \urcorner \wedge \mathsf{T}\dot{\rho}t)$.

For $\varphi = \neg\mathsf{P}_{\geqslant}(t,s)$, if $s^{\mathbb{N}} \notin \mathsf{Rat}$,

$$
(w, f_\alpha^{\mathsf{P}}) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\mathsf{P}_{\geqslant}(t,s)
$$

and

$$
f_\alpha^{\mathsf{P}} \models_{\mathfrak{M}}^{\mathsf{SKT}} \neg(s \preceq \ulcorner 0 \urcorner \vee (s \succ \ulcorner 0 \urcorner \wedge \mathsf{T}\dot{\rho}t)).
$$

For $s^{\mathbb{N}} \preceq 0$,

$$
(w, f_\alpha^{\mathsf{P}}) \not\models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\mathsf{P}_{\geqslant}(t,s)
$$

and

$$
f_\alpha^{\mathsf{P}} \not\models_{\mathfrak{M}}^{\mathsf{SKT}} \neg(s \preceq \ulcorner 0 \urcorner \vee (s \succ \ulcorner 0 \urcorner \wedge \mathsf{T}\dot{\rho}t)).
$$

For $s \succ 0$,

$$
\begin{aligned}
(w, f_\alpha^{\mathsf{P}}) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg\mathsf{P}_{\geqslant}(t,s) \quad &\Longleftrightarrow \quad m_w\{v \mid \dot{\neg}t^{\mathbb{N}} \in f_\alpha^{\mathsf{P}}(v)\} > 1 - \mathsf{rat}(s^{\mathbb{N}}) \\
&\Longleftrightarrow \quad \dot{\neg}t^{\mathbb{N}} \in f_\alpha^{\mathsf{P}}(w) \\
&\Longleftrightarrow \quad \dot{\neg}\rho t^{\mathbb{N}} \in f_\alpha^{\mathsf{T}} \\
&\Longleftrightarrow \quad f_\alpha^{\mathsf{P}} \models_{\mathfrak{M}}^{\mathsf{SKT}} \neg\mathsf{T}\dot{\rho}t \\
&\Longleftrightarrow \quad f_\alpha^{\mathsf{P}} \models_{\mathfrak{M}}^{\mathsf{SKT}} \neg(s \preceq \ulcorner 0 \urcorner \vee (s \succ \ulcorner 0 \urcorner \wedge \mathsf{T}\dot{\rho}t))
\end{aligned}
$$

The inductive steps are clear. $\qquad\square$

Not all sentences are assigned either probability $\{0\}, \{1\}, [0,1]$. We now give an example of such a situation. Consider an agent's degrees of belief in the toss of a fair coin.
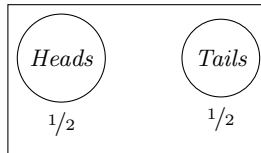

Figure 3.5: Agent considering the toss of a fair coin.

For f the minimal fixed point, at w_{Heads}, $\mathrm{p}(Heads \vee \lambda) = [0.5, 1]$.

80

3.4.4 Other special cases

Question 3.4.8. [18] *What happens in the pms* $\mathfrak{M} = \langle \mathbb{N}, \{m_n\}, \mathbf{M} \rangle$ *where*

$$m_n(\{i\}) = \begin{cases} \frac{1}{n} & i \leqslant n \\ 0 & otherwise \end{cases}$$

Answer. Define $f(w)(\#\varphi) := \begin{cases} 1 & \#\varphi \in f(w) \\ 0 & \text{otherwise} \end{cases}$

$$(n, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_{\geqslant}(\ulcorner \varphi \urcorner, \ulcorner r \urcorner) \iff \frac{f(0)(\#\varphi) + \ldots + f(n)(\#\varphi)}{n} \geqslant r$$

$$(n, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg \mathsf{P}_{\geqslant}(\ulcorner \varphi \urcorner, \ulcorner r \urcorner) \iff \frac{f(0)(\#\neg\varphi) + \ldots + f(n)(\#\neg\varphi)}{n} > 1 - r$$

3.5 An axiomatic system

In the last section of this chapter we present an axiomatic theory for this semantic construction. This will allow one to better reason about this semantics.

3.5.1 The system and a statement of the result

We now present the axiomatic system.

Definition 3.5.1. Remember we introduced the following abbreviations:

- $\mathsf{P}_{>}(t, s) := \exists a \succ s(\mathsf{P}_{\geqslant}(t, a))$

- $\mathsf{P}_{\leqslant}(t, s) := \mathsf{P}_{\geqslant}(\dot{\neg}t, 1\dot{-}s)$

- $\mathsf{P}_{<}(t, s) := \mathsf{P}_{>}(\dot{\neg}t, 1\dot{-}s)$

- $\mathsf{P}_{=}(t, s) := \mathsf{P}_{\geqslant}(t, s) \wedge \mathsf{P}_{\leqslant}(t, s)$

Define ProbKF to be given by the following axioms, added to an axiomatisation of classical logic.[19]

- KF, the axioms for truth:

    1 $\mathsf{PA}^{\mathcal{L}_{\mathsf{P}_{\geqslant}, \mathsf{T}}}$ the axioms of Peano Arithmetic with the induction schema extended to $\mathcal{L}_{\mathsf{P}_{\geqslant}, \mathsf{T}}$.

    2 $\forall x, y \in \mathsf{cClTerm}_{L_{\mathsf{PA}}}(\mathsf{T}y\dot{=}x \leftrightarrow y^\circ = x^\circ)$

    3 $\forall x, y \in \mathsf{cClTerm}_{L_{\mathsf{PA}}}(\mathsf{T}\dot{\neg}y\dot{=}x \leftrightarrow \neg y^\circ = x^\circ)$

    4 $\forall x_1 \ldots x_n \in \mathsf{cClTerm}_{L_{\mathsf{PA}}}(\mathsf{T}\dot{Q}x_1 \ldots x_n \leftrightarrow Qx_1^\circ \ldots x_n^\circ)$ for each n-ary predicate Q of $\mathcal{L}$

    5 $\forall x_1 \ldots x_n \in \mathsf{cClTerm}_{L_{\mathsf{PA}}}(\mathsf{T}\dot{\neg}\dot{Q}x_1 \ldots x_n \leftrightarrow \neg Qx_1^\circ \ldots x_n^\circ)$ for each n-ary predicate Q of $\mathcal{L}$

    6 $\forall x(\mathsf{Sent}_{\mathsf{P}_{\geqslant}, \mathsf{T}}(x) \rightarrow (\mathsf{T}\dot{\neg}\dot{\neg}x \leftrightarrow \mathsf{T}x)$

---

[18]I was asked this question by Stanislav Speranski

[19]In particular $\vdash_{\mathsf{ProbKF}}$, and later $\vdash_{\mathsf{ProbKF} \cup \Sigma}^{\omega}$, should be a classical deducibility relation in the sense of Goldblatt (2014).

7 $\forall x(\mathsf{Sent}_{\mathsf{P}_{\geqslant},\mathsf{T}}(x\dot\vee y) \to (\mathsf{T}x\dot\vee y \leftrightarrow (\mathsf{T}x \vee \mathsf{T}y))$

8 $\forall x(\mathsf{Sent}_{\mathsf{P}_{\geqslant},\mathsf{T}}(x\dot\vee y) \to (\mathsf{T}\dot\neg x\dot\vee y \leftrightarrow (\mathsf{T}\dot\neg x \wedge \mathsf{T}\dot\neg y))$

9 $\forall x(\mathsf{Sent}_{\mathsf{P}_{\geqslant},\mathsf{T}}(\dot\exists vx) \to (\mathsf{T}\dot\exists vx \leftrightarrow \exists y(x(y/v))))$

10 $\forall x(\mathsf{Sent}_{\mathsf{P}_{\geqslant},\mathsf{T}}(\dot\exists vx) \to (\mathsf{T}\dot\neg\dot\exists vx \leftrightarrow \forall y(\dot\neg x(y/v))))$

11 $\forall x \in \mathsf{cClTerm}_{L_{\mathsf{PA}}}(\mathsf{T}\dot{\mathsf{T}}x \leftrightarrow \mathsf{T}x^\circ)$

12 $\forall x \in \mathsf{cClTerm}_{L_{\mathsf{PA}}}(\mathsf{T}\dot\neg\dot{\mathsf{T}}x \leftrightarrow (\mathsf{T}\dot\neg x^\circ \vee \neg\mathsf{Sent}_{\mathsf{P}_{\geqslant},\mathsf{T}}(x^\circ)))$

13 $\forall x(\mathsf{T}x \to \mathsf{Sent}_{\mathsf{P}_{\geqslant},\mathsf{T}}(x))$

- InteractionAx, the axioms for the interaction of truth and probability:[20]

  14 $\forall x,y \in \mathsf{cClTerm}_{L_{\mathsf{PA}}}(\mathsf{TP}_{\geqslant}(x,y) \leftrightarrow \mathsf{P}_{\geqslant}(x^\circ,y^\circ)))$

  15 $\forall x,y \in \mathsf{cClTerm}_{L_{\mathsf{PA}}}(\mathsf{T}\dot\neg\mathsf{P}_{\geqslant}(x,y) \leftrightarrow (\mathsf{P}_{<}(x^\circ,y^\circ) \vee \neg\mathsf{Rat}(y^\circ))))$

- The axioms which give basic facts about $\mathsf{P}_{\geqslant}$:

  16 $\forall a(\exists x\mathsf{P}_{\geqslant}(x,a) \to \mathsf{Rat}a)$

  17 $\forall x(\mathsf{P}_{>}(x,\ulcorner 0\urcorner) \to \mathsf{Sent}_{\mathsf{P}_{\geqslant},\mathsf{T}}x)$

  18 $\forall x\forall a \in \mathsf{Rat}(\mathsf{P}_{\geqslant}(x,a) \leftrightarrow \forall b \prec a\mathsf{P}_{\geqslant}(x,b))$

- Axioms and a rule which say that $\mathsf{P}_{\geqslant r}$ acts like a probability[21]:

  19 $\mathsf{P}_{\geqslant}(\ulcorner 0=0\urcorner,\ulcorner 1\urcorner) \wedge \neg\mathsf{P}_{>}(\ulcorner 0=0\urcorner,\ulcorner 1\urcorner)$

  20 $\mathsf{P}_{\geqslant}(\ulcorner\neg 0=0\urcorner,\ulcorner 0\urcorner) \wedge \neg\mathsf{P}_{>}(\ulcorner\neg 0=0\urcorner,\ulcorner 0\urcorner)$

  21 $\forall x\forall y(\mathsf{Sent}_{\mathsf{P}_{\geqslant},\mathsf{T}}(x) \wedge \mathsf{Sent}_{\mathsf{P}_{\geqslant},\mathsf{T}}(y) \to$
  $\left(\forall a \in \mathsf{Rat}\left(\begin{array}{l}(\forall b\forall c(\quad \mathsf{P}_{\geqslant}(x,b) \quad \wedge \quad \mathsf{P}_{\geqslant}(y,c) \quad \to b\dot+c \preceq a\,)) \\ \leftrightarrow (\forall d\forall e(\,\mathsf{P}_{\geqslant}(x\dot\wedge y,d) \wedge \mathsf{P}_{\geqslant}(x\dot\vee y,e) \to d\dot+e \preceq a\,))\end{array}\right)\right)\right)$

  22 $\dfrac{\mathsf{T}t \to \mathsf{T}s}{\forall a(\mathsf{P}_{\geqslant}(t,a) \to \mathsf{P}_{\geqslant}(s,a))}$

We say $\Gamma \vdash_{\mathsf{ProbKF}} \varphi$ if Rule 22 is used *before* any members of Γ are used.[22]

These axioms are sound, i.e. all induced models satisfy the axiomatisation.

Theorem 3.5.2 (Soundness of ProbKF)**.** *Let* $\mathfrak{M}$ *be a probabilistic structure,* f *a fixed point and* $w \in W$, *and suppose* $\Gamma \vdash_{\mathsf{ProbKF}} \varphi$, *then*

$$\mathsf{IM}_{\mathfrak{M}}[w,f] \models \Gamma \implies \mathsf{IM}_{\mathfrak{M}}[w,f] \models \varphi.$$

Proof. By induction on the length of the proof in ProbKF. Most of the axioms follow from Definition 3.2.3 using the fact that since f is a fixed point $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}\ulcorner\varphi\urcorner \iff (w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \varphi$. □

---

[20]These should be seen as the appropriate way of extending KF to the language $\mathcal{L}_{\mathsf{P}_{\geqslant},\mathsf{T}}$, but we include them separately to highlight them.

[21]We use the axioms for 2-additive Choquet capacities because our underlying structure might be a lattice not a Boolean algebra.

[22]Further details of this constraint can be found in Definition 3.5.3.

We would additionally like to have a completeness component to the axiomatisation. To do this we will expand the result from Feferman that $(\mathbb{N}, S) \models \mathsf{KF}$ iff S is a fixed point. This result was extended to the modal case by Stern (2014b), where Stern shows that KF extended by axioms for the interaction of truth with a necessity and possibility predicate analogous to Axioms 14 and 15 allows one to pick out the fixed points. Lemma 3.5.30 is a minor modifications of Stern's result.

The adequacy of KF relies on the fact that we work with a standard model of arithmetic. So, for us to use this theorem we need to ensure that arithmetic is fixed. To do this we add an ω-rule to the axiomatisation. This allows one to conclude $\forall x \varphi(x)$ from all the instances of $\varphi(\overline{n})$.

In doing this we have to be a little careful about contexts because the ω-rule can be used with non-empty context, whereas the rule, 22, can only be used with empty context. The ω rule is therefore treated like a rule like Modes Ponens. It really acts like an axiom but we cannot state it as an axiom simply because our language is finitary so we don't have the syntactic resources to do so.

So, more carefully, the axiomatic system is as follows:

Definition 3.5.3. The ω-rule is:

$$\frac{\varphi(\overline{0}) \qquad \varphi(\overline{1}) \qquad \ldots}{\forall x \varphi(x)}$$

We say $\Gamma \vdash^{\omega}_{\mathsf{ProbKF} \cup \Sigma} \varphi$ if there is a derivation using ProbKF, Σ and the ω-rule, where Rule 22 and any rules from Σ are used before any members of Γ.[23] I.e. Γ are local premises, Σ and ProbKF are global premises.

More carefully, we say that $\Gamma \vdash^{\omega}_{\mathsf{ProbKF} \cup \Sigma} \varphi$ iff there is sequence of formulas (that is possibly transfinitely long) which is split into two parts, a "global part" and a "local part", satisfying the following criteria:
In either part of the proof, the lines of the proof can:

- Be an axiom of classical logic,

- Follow from all preceding lines by a rule of classical logic,

- Follow from all preceding lines by an application of the ω-rule.

  - More explicitly, the line can be $\forall x \varphi(x)$, if each $\varphi(\overline{n})$ appears somewhere beforehand,

In addition, in the first, global, part of the proof, lines can:

- Be an axiom from ProbKF or Σ,

- Follow from all preceding lines by a rule from ProbKF,

  - More explicitly, the line can be $\forall a(\mathsf{P}_{\geqslant}(t, a) \to \mathsf{P}_{\geqslant}(s, a))$, if $\mathsf{T}t \to \mathsf{T}s$ appears somewhere beforehand,

- Follow from the preceding lines by a rule from Σ.

In addition to the axioms and rules that can be used anywhere, in the second, local part of the proof, lines can:

---

[23]But note that the ω-rule can be used anywhere.

- Be an axiom from Γ,

- Follow from all preceding lines by a rule from Γ.

This result is proved by a canonical model construction. The fact that we can produce a canonical model is independently interesting since it gives a systematic structure which one can use when working with with these semantics. The definition of the canonical model can be found on Page 88 and will be such that:

- W_c^Σ is the set of w such that w is a maximally finitely $\vdash^\omega_{\mathsf{ProbKF} \cup \Sigma}$-consistent set of formulas that is closed under the ω-rule.

- $\mathbf{M}(w)$ is an $\mathbb{N}$-model of $\mathcal{L}$ with $\mathbf{M}(w) \models \varphi$ iff $\varphi \in w$, for $\varphi \in \mathsf{Sent}_\mathcal{L}$.

- $f_c^\Sigma(w) := \{n \mid \mathsf{T}\overline{n} \in w\}$.

- For each $w \in W_c^\Sigma$, $m_w : \wp(W_c^\Sigma) \to \mathbb{R}$ is probabilistic such that

$$m_w(\{v \in W_c^\Sigma \mid \mathsf{T}\overline{n} \in v\}) = \sup\{r \mid \mathsf{P}_\geqslant(\overline{n}, \ulcorner r \urcorner) \in w\}.$$

Σ is being used as global premises, so for a probabilistic modal structure, with evaluation function, to satisfy Σ requires that the model at every world satisfies Σ.

Definition 3.5.4. We say $\mathsf{IM}_\mathfrak{M}[f] \models \Sigma$, where Σ is a collection of rules and axioms, if:

- For each axiom φ in Σ, for each $w \in W$, $\mathsf{IM}_\mathfrak{M}[w, f] \models \varphi$, and

- For each rule $\Lambda \rightsquigarrow \varphi$ in Σ, if for every $w \in W$, $\mathsf{IM}_\mathfrak{M}[w, f] \models \Lambda$, then for each $w \in W$, $\mathsf{IM}_\mathfrak{M}[w, f] \models \varphi$.

We will then show:

Theorem 3.5.5 (Soundness and Completeness)**.** *The following are equivalent:*

1. $\Gamma \vdash^\omega_{\mathsf{ProbKF} \cup \Sigma} \varphi$,

2. *For every $w \in W_c^\Sigma$, $\mathsf{IM}_{\mathfrak{M}_c^\Sigma}[w, f_c^\Sigma] \models \Gamma \implies \mathsf{IM}_{\mathfrak{M}_c^\Sigma}[w, f_c^\Sigma] \models \varphi$.*

3. *For every probabilistic modal structure $\mathfrak{M}$ with fixed point f, such that $\mathsf{IM}_\mathfrak{M}[f] \models \Sigma$, we have that for each $w \in W$,*

$$\mathsf{IM}_\mathfrak{M}[w, f] \models \Gamma \implies \mathsf{IM}_\mathfrak{M}[w, f] \models \varphi.$$

We will also have the following forms of the result:

Theorem 3.5.6. *Let $\mathcal{M}$ be a $\mathcal{L}_{\mathsf{P}_\geqslant, \mathsf{T}}$-model.*

1. *The following are equivalent:*

    (a) $\mathcal{M} \models \Gamma \implies \mathcal{M} \models \varphi$ *whenever* $\Gamma \vdash^\omega_{\mathsf{ProbKF} \cup \Sigma} \varphi$,

    (b) $\mathsf{Theory}(\mathcal{M})$ *is* $\vdash^\omega_{\mathsf{ProbKF} \cup \Sigma}$-*consistent,*

(c) *There is a probabilistic structure $\mathfrak{M}$ and fixed point f where $\mathsf{IM}_{\mathfrak{M}}[f] \models \Sigma$, and there is some $w \in W$ such that $\mathcal{M}$ is elementarily equivalent to $\mathsf{IM}_{\mathfrak{M}}[w, f]$.[24]*

2. *Suppose $\mathcal{M}$ is an $\mathbb{N}$-model.[25] Then the following are equivalent:*

   (a) *$\mathcal{M} \models \varphi$ for each $\vdash^{\omega}_{\mathsf{ProbKF}} \varphi$,*

   (b) *$\mathsf{Theory}(\mathcal{M})$ is finitely $\vdash^{\omega}_{\mathsf{ProbKF} \cup \Sigma}$-consistent,*

   (c) *There is a probabilistic structure $\mathfrak{M}$ and fixed point f where $\mathsf{IM}_{\mathfrak{M}}[f] \models \Sigma$, and there is some $w \in W$ with $\mathcal{M} = \mathsf{IM}_{\mathfrak{M}}[w, f]$.*

3.5.2 Proof of the soundness and completeness of ProbKF^{ω}

We first mention a lemma and then work through the soundness and then the completeness results.

Lemma 3.5.7. *Fix $\mathfrak{M}$ a probabilistic modal structure. f is a fixed point iff each $f(w) \subseteq \mathsf{Sent}_{\mathsf{P}_{\geqslant}, \mathsf{T}}$ and for every φ,*

$$\mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{T}\ulcorner\varphi\urcorner \iff (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \varphi$$

Proof. Suppose each $f(w) \subseteq \mathsf{Sent}_{\mathsf{P}_{\geqslant}, \mathsf{T}}$.

$$\#\varphi \in f(w) \iff (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{T}\ulcorner\varphi\urcorner \iff \mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{T}\ulcorner\varphi\urcorner$$

$$\#\varphi \in \Theta(f)(w) \iff (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \varphi$$

Therefore each

$$\mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{T}\ulcorner\varphi\urcorner \iff (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \varphi$$

holds iff f is a fixed point.

Suppose $f(w) \not\subseteq \mathsf{Sent}_{\mathsf{P}_{\geqslant}, \mathsf{T}}$. By construction $\Theta(f)(w) \subseteq \mathsf{Sent}_{\mathsf{P}_{\geqslant}, \mathsf{T}}$, therefore $\Theta(f)(w) \neq f(w)$, i.e. f is not a fixed point. $\qquad\square$

Soundness

Theorem 3.5.8 (Soundness). *Suppose $\Gamma \vdash^{\omega}_{\mathsf{ProbKF} \cup \Sigma} \varphi$. Let $\mathfrak{M}$ be a probabilistic structure, f a fixed point such that $\mathsf{IM}_{\mathfrak{M}}[f] \models \Sigma$. Then for each $w \in W$:*

$$\mathsf{IM}_{\mathfrak{M}}[w, f] \models \Gamma \implies \mathsf{IM}_{\mathfrak{M}}[w, f] \models \varphi.$$

Proof. Fix $\mathfrak{M}$ a probabilistic modal structure and f a fixed point such that $\mathsf{IM}_{\mathfrak{M}}[f] \models \Sigma$. Work by transfinite induction on the depth of the proof of $\Gamma \vdash^{\omega}_{\mathsf{ProbKF} \cup \Sigma} \varphi$ to show that for each $w \in W$,

$$\mathsf{IM}_{\mathfrak{M}}[w, f] \models \Gamma \implies \mathsf{IM}_{\mathfrak{M}}[w, f] \models \varphi$$

---

[24]I.e. $\mathcal{M}$ and $\mathsf{IM}_{\mathfrak{M}}[w, f]$ satisfy all the same $\mathcal{L}_{\mathsf{P}_{\geqslant}, \mathsf{T}}$-sentences.

[25]We still need this assumption because by assumption all $\mathsf{IM}_{\mathfrak{M}}[w, f]$ are $\mathbb{N}$-models, but even adding the ω-rule does not fix the standard model of arithmetic, it only fixes the *theory* of the standard model of arithmetic.

The base case are the logical axioms and the axioms of ProbKF and Σ. The logical axioms are satisfied by $\mathsf{IM}_{\mathfrak{M}}[w, f] \models wf$ because it is a classical model. The axioms of Σ are satisfied by assumption on $\mathfrak{M}$. We reason for the ProbKF axioms as follows.

Axiom 1 holds because $\mathsf{IM}_{\mathfrak{M}}[w, f]$ is an $\mathbb{N}$-model by definition.

The KF axioms Axioms 2 to 10 can be seen as directly following from the semantic clauses of Definition 3.2.3 using Lemma 3.5.7. For example:

- Axiom 7:
$$\mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{T}^\ulcorner \varphi \vee \psi \urcorner$$

iff	$(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \varphi \vee \psi$	Lemma 3.5.7
iff	$(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \varphi$ or $(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \psi$	Definition 3.2.3
iff	$\mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{T}^\ulcorner \varphi \urcorner \vee \mathsf{T}^\ulcorner \psi \urcorner$	Lemma 3.5.7

  Therefore: $\mathsf{IM}_{\mathfrak{M}}[w, f] \models \forall x, y(\mathsf{Sent}_{\mathsf{P}_\geqslant, \mathsf{T}}(x \dot{\vee} y) \to (\mathsf{T} x \dot{\vee} y \leftrightarrow (\mathsf{T} x \vee \mathsf{T} y))$

- Axiom 10:
$$\mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{T}^\ulcorner \neg \exists x \varphi(x) \urcorner$$

iff	$(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg \exists x \varphi(x)$	Lemma 3.5.7
iff	$(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg \varphi(\overline{n})$ for all n	Definition 3.2.3
iff	$\mathsf{IM}_{\mathfrak{M}}[w, f] \models \neg \varphi(\overline{n})$ for all n	Lemma 3.5.7
iff	$\mathsf{IM}_{\mathfrak{M}}[w, f] \models \neg \exists x \varphi(x)$	

Axiom 13: By Lemma 3.5.7, $f(w) \subseteq \mathsf{Sent}_{\mathsf{P}_\geqslant, \mathsf{T}}$. Therefore $\mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{T} t \implies t^{\mathbb{N}} \in \mathsf{Sent}_{\mathsf{P}_\geqslant, \mathsf{T}}$.

Axioms 11, 12, 14 and 15 follow directly from Lemma 3.5.7, Definition 3.2.12, and Proposition 3.2.4. For example:

- Axiom 11:
$$\mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{T}^\ulcorner \mathsf{T} t \urcorner$$

iff	$(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{T} t$	Lemma 3.5.7
iff	$\mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{T} t$	Definition 3.2.12

- Axiom 15:
$$\mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{T}^\ulcorner \neg \mathsf{P}_\geqslant(t, s) \urcorner$$

iff	$(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \neg \mathsf{P}_\geqslant(t, s)$	Lemma 3.5.7
iff	$(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_>(\dot{\neg} t, 1 \dot{-} s)$ or $s^{\mathbb{N}} \notin \mathsf{Rat}$	Proposition 3.2.4
iff	$\mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{P}_>(\dot{\neg} t, 1 \dot{-} s) \vee \neg \mathsf{Rat}(s)$	Definition 3.2.12

For Axioms 16 to 21 and Rule 22, individual arguments need to be given.

- Axiom 16:
$$\mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{P}_\geqslant(\overline{n}, \overline{k})$$

$\Longleftrightarrow$	$(w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_\geqslant(\overline{n}, \overline{k})$	Definition 3.2.12
$\Longrightarrow$	$k \in \mathsf{Rat}$	Definition 3.2.3

  Therefore for each $k, n \in \mathbb{N}$, $\mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{P}_\geqslant(\overline{n}, \overline{k}) \to \mathsf{Rat}(\overline{k})$ and so $\mathsf{IM}_{\mathfrak{M}}[w, f] \models \forall a(\exists x \mathsf{P}_\geqslant(x, a) \to \mathsf{Rat}(a))$ since $\mathsf{IM}_{\mathfrak{M}}[w, f]$ is a standard model.

- Axiom 17:
$$\mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{P}_>(\overline{n}, ^\ulcorner 0 \urcorner)$$

$\Longleftrightarrow$	$m_w\{v \mid n \in f(v)\} > 0$	Definitions 3.2.3 and 3.2.12
$\Longrightarrow$	$\{v \mid n \in f(v)\} \neq \varnothing$	
$\Longrightarrow$	$n \in \mathsf{Sent}_{\mathsf{P}_\geqslant, \mathsf{T}}$	Lemma 3.5.7 and that f is a fixed point.

- Axiom 18: We prove the "→" and "←" separately.
  First "→":
  $$\mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{P}_{\geqslant}(\overline{n}, \ulcorner r \urcorner) \text{ and } r > q$$
$\implies$	$m_w\{v \mid n \in f(v)\} \geqslant r > q$	Definitions 3.2.3 and 3.2.12
$\implies$	$\mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{P}_{\geqslant}(\overline{n}, \ulcorner q \urcorner)$	Definitions 3.2.3 and 3.2.12

  Since $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \forall b(b \prec \ulcorner r \urcorner \to \mathsf{Rat}(b))$ this suffices.
  Now for "←":
  $$\mathsf{IM}_{\mathfrak{M}}[w,f] \models \forall b \prec \ulcorner r \urcorner \mathsf{P}_{\geqslant}(\overline{n}, b)$$
$\implies$	for all rationals $q < r$, $\quad m_w\{v \mid n \in f(v)\} \geqslant q$	Definitions 3.2.3 and 3.2.12
$\implies$	$m_w\{v \mid n \in f(v)\} \geqslant r$	$\mathbb{Q}$ is dense in $\mathbb{R}$
$\implies$	$\mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{P}_{\geqslant}(\overline{n}, \ulcorner r \urcorner)$	Definitions 3.2.3 and 3.2.12

- Axiom 19
	For each $v \in W$, $(v,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} 0 = 0$	Definition 3.2.3, and $\mathbf{M}(v)$ is an $\mathbb{N}$-model
so	For each $v \in W$, $\#0 = 0 \in f(v)$	Proposition 3.2.11
so	$m_w\{v \mid \#0 = 0 \in f(v)\} = m_w(W) = 1$	

  Therefore $(w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{\geqslant}(\ulcorner 0 = 0 \urcorner, \ulcorner 1 \urcorner)$ using this and Definition 3.2.3. Also for each $r > 1$, $(w,f) \not\models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{\geqslant}(\ulcorner 0 = 0 \urcorner, \ulcorner r \urcorner)$. By using Definition 3.2.12 and the definition of $\mathsf{P}_{>}(t,s)$ we have that $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{P}_{\geqslant}(\ulcorner 0 = 0 \urcorner, \ulcorner 1 \urcorner) \wedge \neg \mathsf{P}_{>}(\ulcorner 0 = 0 \urcorner, \ulcorner 1 \urcorner)$.

- Axiom 20 The argument is analogous to that of Axiom 19 but we now show that $m_w\{v \mid \#\neg 0 = 0 \in f(v)\} = m_w(\varnothing) = 0$.

- Axiom 21:
  For $\varphi \in \mathsf{Sent}_{\mathsf{P}_{\geqslant}, \mathsf{T}}$ let
  $$[\varphi] := \{v \mid (w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \varphi\}.$$

  By using Definitions 3.2.3 and 3.2.12, one can see that $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{P}_{\geqslant}(\ulcorner \varphi \urcorner, \ulcorner r \urcorner) \iff m_w[\varphi] \geqslant r$.

  By Definition 3.2.3,

  $$[\varphi] \cup [\psi] = [\varphi \vee \psi] \text{ and } [\varphi] \cap [\psi] = [\varphi \wedge \psi]$$

  so $m_w[\varphi] + m_w[\psi] = m_w[\varphi \vee \psi] + m_w[\varphi \wedge \psi]$

  so $\forall r \in \mathbb{Q}, (m_w[\varphi] + m_w[\psi] \leqslant r \iff m_w[\varphi \vee \psi] + m_w[\varphi \wedge \psi] \leqslant r)$

  Now,

  $$m_w[\varphi] + m_w[\psi] \leqslant r$$
  $$\iff \forall q, \gamma \in \mathbb{Q}(m_w[\varphi] \geqslant q \wedge m_w[\psi] \geqslant \gamma \to q + \gamma \leqslant r)$$
  $$\iff \mathsf{IM}_{\mathfrak{M}}[w,f] \models \forall b, c(\mathsf{P}_{\geqslant}(\ulcorner \varphi \urcorner, b) \wedge \mathsf{P}_{\geqslant}(\ulcorner \psi \urcorner, c) \to b \dotplus c \preceq \ulcorner r \urcorner)$$

  Similarly,

  $$m_w[\varphi \wedge \psi] + m_w[\varphi \vee \psi] \leqslant r$$
  $$\iff \mathsf{IM}_{\mathfrak{M}}[w,f] \models \forall d, e(\mathsf{P}_{\geqslant}(\ulcorner \varphi \wedge \psi \urcorner, d) \wedge \mathsf{P}_{\geqslant}(\ulcorner \varphi \vee \psi \urcorner, e) \to d \dotplus e \preceq \ulcorner r \urcorner)$$

  Therefore,

  $$\mathsf{IM}_{\mathfrak{M}}[w,f] \models (\forall b, c(\mathsf{P}_{\geqslant}(\ulcorner \varphi \urcorner, b) \wedge \mathsf{P}_{\geqslant}(\ulcorner \psi \urcorner, c) \to b \dotplus c \preceq \ulcorner r \urcorner)$$
  $$\leftrightarrow (\forall d, e(\mathsf{P}_{\geqslant}(\ulcorner \varphi \wedge \psi \urcorner, d) \wedge \mathsf{P}_{\geqslant}(\ulcorner \varphi \vee \psi \urcorner, e) \to d \dotplus e \preceq \ulcorner r \urcorner)$$

That suffices to prove the base case. Now for the induction step: for this we need to consider the logical rules,[26] Rule 22 and the ω-rule.

The logical rules hold because we are only working with classical models.

For Rule 22: Suppose the last step of $\Gamma \vdash^{\omega}_{\mathsf{ProbKF} \cup \Sigma} \varphi$ was an application of Rule 22. Then it must in fact be that $\Gamma = \varnothing$ by Definition 3.5.3. So t and s must therefore be such that

$$\vdash^{\omega}_{\mathsf{ProbKF} \cup \Sigma} \mathsf{T}t \to \mathsf{T}s.$$

Then using the induction hypothesis it must be that for all $v \in W$, $\mathsf{IM}_{\mathfrak{M}}[v, f] \models \mathsf{T}t \to \mathsf{T}s$. But therefore

$$\{v \mid s^{\mathbb{N}} \in f(v)\} \subseteq \{v \mid t^{\mathbb{N}} \in f(v)\}$$

by Definitions 3.2.3 and 3.2.12. So

$$m_w\{v \mid s^{\mathbb{N}} \in f(v)\} \geqslant r \implies m_w\{v \mid t^{\mathbb{N}} \in f(v)\} \geqslant r.$$

So

$$(w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{\geqslant}(s, \ulcorner r \urcorner) \implies (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{\geqslant}(t, \ulcorner r \urcorner)$$

And therefore

$$\mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{P}_{\geqslant}(s, \ulcorner r \urcorner) \to \mathsf{P}_{\geqslant}(t, \ulcorner r \urcorner).$$

Lastly we consider the ω-rule. Suppose the last step of $\Gamma \vdash^{\omega}_{\mathsf{ProbKF} \cup \Sigma} \varphi$ was an application of the ω-rule. Then we must have that $\varphi = \forall x \psi(x)$ and that for each n, $\Gamma \vdash^{\omega}_{\mathsf{ProbKF} \cup \Sigma} \psi(\overline{n})$ by a shorter proof. So by the induction hypothesis we have that $\mathsf{IM}_{\mathfrak{M}}[w, f] \models \Gamma \implies \mathsf{IM}_{\mathfrak{M}}[w, f] \models \psi(\overline{n})$ for each n. Now suppose that in fact $\mathsf{IM}_{\mathfrak{M}}[w, f] \models \Gamma$. Then because $\mathsf{IM}_{\mathfrak{M}}[w, f]$ is an $\mathbb{N}$-model we have that $\mathsf{IM}_{\mathfrak{M}}[w, f] \models \forall x \psi(x)$. $\qquad \square$

We can now turn to the more interesting completeness direction. We prove this by a canonical model construction.

Definition of the canonical model and showing it is well-defined

Definition 3.5.9. For each Σ, define a probabilistic structure $\mathfrak{M}^{\Sigma}_{\mathrm{c}}$ and evaluation function f^{Σ}_{c} as follows:

- W^{Σ}_{c} is the set of w such that w is a maximally finitely $\vdash^{\omega}_{\mathsf{ProbKF} \cup \Sigma}$-consistent set of formulas[27] that is closed under the ω-rule.[28][29]

- Define each $\mathbf{M}(w)$ as an $\mathbb{N}$-model of $\mathcal{L}$ with $\mathbf{M}(w) \models \varphi$ iff $\varphi \in w$, for $\varphi \in \mathsf{Sent}_{\mathcal{L}}$.

- $f^{\Sigma}_{\mathrm{c}}(w) := \{n \mid \mathsf{T}\overline{n} \in w\}$.

---

[26] These would usually be modes ponens and $\forall$-generalisation.

[27] I.e. there is no finite $\Delta \subseteq w$ with $\Delta \vdash^{\omega}_{\mathsf{ProbKF} \cup \Sigma} \bot$

[28] I.e. whenever $\{\varphi(\overline{n}) \mid n \in \mathbb{N}\} \subseteq w$ then $\forall x \varphi(x) \in w$.

[29] In fact such w are exactly the maximally $\vdash^{\omega}_{\mathsf{ProbKF} \cup \Sigma}$-consistent set of formulas. But proving that would take us too far afield. It can be seen as a corollary of Lemma 3.5.29 and Theorem 3.5.8.

- For each $w \in W_c^\Sigma$, find $m_w : \wp(W_c^\Sigma) \to \mathbb{R}$ probabilistic such that

$$m_w(\{v \in W_c^\Sigma \mid \mathsf{T}\overline{n} \in v\}) = \sup\{r \mid \mathsf{P}_{\geqslant}(\overline{n}, \ulcorner r \urcorner) \in w\}.$$

We will now show that this canonical model is well defined. To do this we need to show that we can pick such an $\mathbf{M}(w)$ and m_w. We do that in Lemma 3.5.11 and Theorem 3.5.28.

We first state some facts about $\vdash^\omega_{\mathsf{ProbKF} \cup \Sigma}$ and the $w \in W_c^\Sigma$ that we will use in our proof without further comment.

Lemma 3.5.10. $\vdash^\omega_{\mathsf{ProbKF} \cup \Sigma}$ *has the following properties:*

1. *If* $\Gamma \vdash^\omega_{\mathsf{ProbKF} \cup \Sigma} \varphi$ *and* $\Gamma \cup \{\varphi\} \vdash^\omega_{\mathsf{ProbKF} \cup \Sigma} \bot$ *implies* $\Gamma \vdash^\omega_{\mathsf{ProbKF} \cup \Sigma} \bot$

2. $\Gamma \cup \{\neg\varphi\}^{30}$ *is* $\vdash^\omega_{\mathsf{ProbKF} \cup \Sigma}$-*inconsistent*[31] *iff* $\Gamma \vdash^\omega_{\mathsf{ProbKF} \cup \Sigma} \varphi$.

3. *If* Γ *is* $\vdash^\omega_{\mathsf{ProbKF} \cup \Sigma}$-*consistent and* $\Gamma \vdash^\omega_{\mathsf{ProbKF} \cup \Sigma} \varphi$ *then* $\Gamma \cup \{\varphi\}$ *is* $\vdash^\omega_{\mathsf{ProbKF} \cup \Sigma}$-*consistent.*

4. *If* Γ *is finitely* $\vdash^\omega_{\mathsf{ProbKF} \cup \Sigma}$-*consistent then so is either* $\Gamma \cup \{\varphi\}$ *or* $\Gamma \cup \{\neg\varphi\}$.

For each $w \in W_c^\Sigma$:

5. $\neg\varphi \in w \iff \varphi \notin w$.

6. *If* $\vdash^\omega_{\mathsf{ProbKF} \cup \Sigma} \varphi$ *then* $\varphi \in w$.

7. $\Gamma \vdash_{\mathsf{cl}} \varphi^{32}$ *and* $\Gamma \subseteq w$ *then* $\varphi \in w$.

8. $\varphi \wedge \psi \in w \iff \varphi \in w$ *and* $\psi \in w$.

9. $\forall x \varphi(x) \in w$ *iff for all* $k \in \mathbb{N}$, $\varphi(\overline{k}) \in w$.

To pick such an $\mathbf{M}(w)$ we could use the result that a theory has an $\mathbb{N}$-model if it is closed under the ω-rule (e.g. Chang and Keisler, 1990, Proposition 2.2.12). We prove the result directly here.

Lemma 3.5.11. *For each* $w \in W_c^\Sigma$, *there is some* $\mathbf{M}(w)$ *an* $\mathbb{N}$-*model of* $\mathcal{L}$ *with* $\mathbf{M}(w) \models \varphi$ *iff* $\varphi \in w$, *for* $\varphi \in \mathsf{Sent}_{\mathcal{L}}$.

Proof. Define $\mathbf{M}(w)$ as follows:

Domain is $\mathbb{N}$. Interprets the arithmetic vocabulary as in the standard model of arithmetic. For Q a contingent n-ary relation symbol take $\langle k_1, \ldots k_n \rangle \in Q^{\mathbf{M}(w)}$ iff $Q(\overline{k_1}, \ldots, \overline{k_n}) \in w$.

We prove by induction on the complexity of φ that $\mathbf{M}(w) \models \varphi$ iff $\varphi \in w$.

For φ atomic: If φ is an atomic sentence of $\mathcal{L}_{\mathsf{PA}}$, the result holds because ProbKF extends PA because of Axiom 1.

$$\mathbf{M}(w) \models \varphi \implies \mathbb{N} \models \varphi \implies \vdash^\omega_{\mathsf{ProbKF} \cup \Sigma} \varphi \implies \varphi \in w.$$

$$\mathbf{M}(w) \not\models \varphi \implies \mathbb{N} \not\models \varphi \implies \vdash^\omega_{\mathsf{ProbKF} \cup \Sigma} \neg\varphi \implies \neg\varphi \in w \implies \varphi \notin w.$$

---

[30] I.e. φ is a classical logical consequence of Γ.
[31] I.e. $\Gamma \vdash^\omega_{\mathsf{ProbKF} \cup \Sigma} \bot$.
[32] Where $\vdash_{\mathsf{cl}}$ denotes derivability in classical logic.

If φ is non-arithmetic, atomic, then $\varphi = Q(t_1, \ldots, t_n)$ with t_i closed terms.

$$\begin{aligned}
\mathbf{M}(w) \models Q(t_1, \ldots, t_n) &\iff \mathbf{M}(w) \models Q(\overline{t_1^{\mathbb{N}}}, \ldots, \overline{t_n^{\mathbb{N}}}) \\
&\iff Q(\overline{t_1^{\mathbb{N}}}, \ldots, \overline{t_n^{\mathbb{N}}}) \in w \\
&\iff Q(t_1, \ldots, t_n) \in w,
\end{aligned}$$

the last equivalence holding because of Item 7.

The induction steps are easy, using Items 5, 8 and 9 $\hfill\square$

The axioms in Axioms 16 to 21 and Rule 22 allow us to find such an m_w. To show this we first present a few lemmas.

The following theorem is already known. It's statement can be found, for example, in Zhou (2013). I am currently unaware of where the theorem is from and what its existing proofs are.

Theorem 3.5.12. *Let $\Xi \subseteq \wp(W)$ be a lattice,[33] i.e. a collection of subsets of W closed under $\cup$ and $\cap$ and containing $\varnothing$ and W. For a set $\Lambda \subseteq \wp(W)$, let $\mathfrak{B}(\Lambda)$ be the Boolean closure of Λ.[34] Let m be a monotone 2-valuation on Ξ, i.e. $m : \Xi \to [0,1]$ satisfying:*

- *$m(W) = 1$, $m(\varnothing) = 0$,*

- *Monotonicity: $A \subseteq C \implies m(A) \leqslant m(C)$*

- *2-valuation: $m(A \cup C) + m(A \cap C) = m(A) + m(C)$*

Then there is a function $m^* : \mathfrak{B}(\Xi) \to \mathbb{R}$ which is a (finitely additive) probability function that extends m.
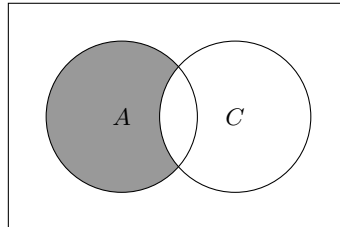
Proof.

Lemma 3.5.13. *If $\{A_i \mid i \in I\}$ is a finite partition of W and $a_i \geqslant 0$ with $\sum_{i \in I} a_i = 1$ then there is a unique probability function $m^* : \mathfrak{B}(\{A_i \mid i \in I\}) \to \mathbb{R}$*

Definition 3.5.14. *We let $\bigcap \varnothing$ denote W. For $\Delta \subseteq \Lambda \subseteq \wp(W)$, define:*

$$E_\Delta^\Lambda := \bigcap \Delta \setminus \bigcup (\Lambda \setminus \Delta).$$

This gives the event which is exactly in all the events of Δ and nothing else. For example $E_{\{A\}}^{\{A,C\}} = A \setminus C$ is the shaded component:

---

[33]Since the powerset algebra is distributive, this is then a distributive lattice.
[34]I.e. the smallest set $\Delta \supseteq \Lambda$ closed under $\cup$, $\cap$ and taking complements, and containing $\varnothing$ and W.

Lemma 3.5.15. $\{E_\Delta^\Lambda \mid \Delta \subseteq \Lambda\}$ *partitions* W. *And each* $E_\Delta^\Lambda \in \mathfrak{B}(\Lambda)$.

Moreover, each member of $\mathfrak{B}(\Lambda)$ is a disjoint union of some of the E_Δ^Λ:

Lemma 3.5.16. *For each* $A \in \mathfrak{B}(\Lambda)$ *and each* E_Δ^Λ, *either* $E_\Delta^\Lambda \subseteq A$, *or* $E_\Delta^\Lambda \subseteq W \setminus A$.

Proof. By induction on the Boolean construction.

Suppose $A \in \Lambda$. We can show: if $A \in \Delta$ then $E_\Delta^\Lambda \subseteq A$, and if $A \notin \Delta$ then $E_\Delta^\Lambda \subseteq W \setminus A$.

Suppose $E_\Delta^\Lambda \not\subseteq A \cup C$. Then it is $\not\subseteq A$, so is $\subseteq W \setminus A$ using the induction hypothesis. And similarly for C. So $E_\Delta^\Lambda \subseteq (W \setminus A) \cap (W \setminus C) = W \setminus (A \cup C)$.

Suppose $E_\Delta^\Lambda \not\subseteq W \setminus A$. Then using the induction hypothesis it is $\subseteq A = W \setminus (W \setminus A)$. $\square$

Corollary 3.5.17. *For each* $A \in \mathfrak{B}(\Lambda)$ $A = \bigcup_{E_\Delta^\Lambda \subseteq A} E_\Delta^\Lambda$

So we just need to pick a_Δ^Λ which will act as the probability of E_Δ^Λ in a way that will give a probability function that extends m.

Definition 3.5.18. Fix $m : \Xi \to \mathbb{R}$ where Ξ is a lattice of subsets of W.

For each $\Lambda \subseteq \Xi$ finite, define:

$$a_\Delta^\Lambda := m\left(\bigcap \Delta\right) - m\left(\bigcap \Delta \cap \bigcup(\Lambda \setminus \Delta)\right)$$

We gave this definition because if m^* is a probability function then

$$m^*(E_\Delta^\Lambda) = m^*\left(\bigcap \Delta \setminus \bigcup(\Lambda \setminus \Delta)\right) = m^*\left(\bigcap \Delta\right) - m^*\left(\bigcap \Delta \cap \bigcup(\Lambda \setminus \Delta)\right)$$

where all the components on the right hand side are in Ξ. So this should get us a probability function that extends m if we choose m^* such that $m^*(E_\Delta^\Lambda) = a_\Delta^\Lambda$.

For the rest of the proof we shall assume that m satisfies:

- $m(W) = 1$, $m(\varnothing) = 0$,
- $A \subseteq C \implies m(A) \leqslant B(C)$
- $m(A \cup C) + m(A \cap C) = m(A) + m(C)$

Observe that for $\Delta \subseteq \Lambda$, $E_\Delta^{\Lambda \cup \{D\}} = E_\Delta^\Lambda \setminus D$ and $E_{\Delta \cup \{D\}}^{\Lambda \cup \{D\}} = E_\Delta^\Lambda \cap D$, so these partition E_Δ^Λ.

For example if $\Lambda = \{A, C\}$ and $\Delta = \{A, C\}$, $E_\Delta^{\Lambda \cup \{D\}}$ is the shaded part, $E_{\Delta \cup \{D\}}^{\Lambda \cup \{D\}}$ is the hashed part, and E_Δ^Λ is the union of these.
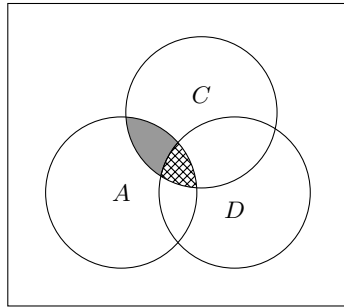
So we would hope that $a_\Delta^{\Lambda \cup \{D\}} + a_{\Delta \cup \{D\}}^{\Lambda \cup \{D\}} = a_\Delta^\Lambda$, which we do in fact find:

Lemma 3.5.19. $a_\Delta^{\Lambda \cup \{D\}} + a_{\Delta \cup \{D\}}^{\Lambda \cup \{D\}} = a_\Delta^\Lambda$

Proof.

$$a_\Delta^{\Lambda \cup \{D\}} = m\left(\bigcap \Delta\right) - m\left(\bigcap \Delta \cap \left(\bigcup(\Lambda \setminus \Delta) \cup D\right)\right) \tag{3.3}$$

$$= m\left(\bigcap \Delta\right) - m\left(\left(\bigcap \Delta \cap \bigcup(\Lambda \setminus \Delta)\right) \cup \left(\bigcap \Delta \cap D\right)\right) \tag{3.4}$$

$$= m\left(\bigcap \Delta\right) - m\left(\left(\bigcap \Delta \cap \bigcup(\Lambda \setminus \Delta)\right)\right) - m\left(\left(\bigcap \Delta \cap D\right)\right)$$
$$+ m\left(\left(\bigcap \Delta \cap \bigcup(\Lambda \setminus \Delta)\right) \cap \left(\bigcap \Delta \cap D\right)\right) \tag{3.5}$$

$$= m\left(\bigcap \Delta\right) - m\left(\left(\bigcap \Delta \cap \bigcup(\Lambda \setminus \Delta)\right)\right) - m\left(\left(\bigcap \Delta \cap D\right)\right)$$
$$+ m\left(\bigcap \Delta \cap D \cap \bigcup(\Lambda \setminus \Delta)\right) \tag{3.6}$$

The step from Eq. (3.3) to Eq. (3.4) just involves rearranging the expression. The step from Eq. (3.4) to Eq. (3.5) works by applying the fact that m is a 2-valuation. Lastly to Eq. (3.6) is again just a rearrangement. We now add this to $a_\Delta^{\Lambda \cup \{D\}}$:

$$a_\Delta^{\Lambda \cup \{D\}} + a_{\Delta \cup \{D\}}^{\Lambda \cup \{D\}} = m\left(\bigcap \Delta\right) - m\left(\left(\bigcap \Delta \cap \bigcup(\Lambda \setminus \Delta)\right)\right)$$
$$- m\left(\left(\bigcap \Delta \cap D\right)\right)$$
$$+ m\left(\left(\bigcap \Delta \cap D \cap \bigcup(\Lambda \setminus \Delta)\right)\right)$$
$$+ m\left(\bigcap \Delta \cap D\right) - m\left(\bigcap \Delta \cap D \cap \bigcup(\Lambda \setminus \Delta)\right)$$
$$= m\left(\bigcap \Delta\right) - m\left(\left(\bigcap \Delta \cap \bigcup(\Lambda \setminus \Delta)\right)\right)$$
$$= a_\Delta^\Lambda$$

This works just by cancelling terms. $\qquad \square$

Lemma 3.5.20. *Each* $a_\Delta^\Lambda \geqslant 0$ *and* $\sum_{\Delta \subseteq \Lambda} a_\Delta^\Lambda = 1$

Proof.

$$\bigcap \Delta \supseteq \bigcap \Delta \cap \left(\bigcup \Lambda \setminus \Delta\right)$$
$$\text{so } m\left(\bigcap \Delta\right) \geqslant m\left(\bigcap \Delta \cap \left(\bigcup \Lambda \setminus \Delta\right)\right)$$
$$\text{so } a_\Delta^\Lambda = m\left(\bigcap \Delta\right) - m\left(\bigcap \Delta \cap \left(\bigcup \Lambda \setminus \Delta\right)\right) \geqslant 0$$

by the monotonicity requirement on m.

For $\sum_{\Delta \subseteq \Lambda} a_\Delta^\Lambda = 1$ we work by induction on the size of Λ. The base case is

clear because for $\Lambda = \{A\}$,

$$\sum_{\Delta \subseteq \{A\}} a_\Delta^\Lambda = a_{\{A\}}^{\{A\}} + a_\varnothing^{\{A\}}$$

$$= m(\bigcap\{A\}) - m\left(\bigcap\{A\} \cap \bigcup(\{A\} \setminus \{A\})\right)$$

$$+ m(\bigcap \varnothing) - m\left(\bigcap \varnothing \cap \bigcup(\{A\} \setminus \varnothing)\right)$$

$$= m(A) - m(\varnothing) + m(W) - m(A)$$

$$= m(W) - m(\varnothing) = 1$$

For the induction step we use Lemma 3.5.19 to show

$$1 = \sum_{\Delta \subseteq \Lambda} a_\Delta^\Lambda = \sum_{\Delta \subseteq \Lambda} (a_\Delta^{\Lambda \cup \{D\}} + a_{\Delta \cup \{D\}}^{\Lambda \cup \{D\}}) = \sum_{\Delta \subseteq \Lambda \cup \{D\}} a_\Delta^{\Lambda \cup \{D\}} \qquad \square$$

Corollary 3.5.21. *For each finite $\Lambda \subseteq \Xi$, there is a unique probability function $m_\Lambda^* : \mathfrak{B}(\Lambda) \to \mathbb{R}$ such that for each $\Delta \subseteq \Lambda$, $m_\Lambda^*\left(E_\Delta^\Lambda\right) = a_\Delta^\Lambda$. This is given by for $A \in \mathfrak{B}(\Lambda)$, $m_\Lambda^*(A) = \sum_{E_\Delta^\Lambda \subseteq A} a_\Delta^\Lambda$.*

Lemma 3.5.22. *If $A \in \mathfrak{B}(\Lambda)$ then*

$$E_\Delta^\Lambda \subseteq A \iff E_\Delta^{\Lambda \cup \{D\}} \subseteq A$$

and

$$E_\Delta^\Lambda \subseteq A \iff E_{\Delta \cup \{D\}}^{\Lambda \cup \{D\}} \subseteq A$$

Proof. For any $\Delta \subseteq \Lambda$, $E_\Delta^{\Lambda \cup \{D\}} = E_\Delta^\Lambda \setminus D$ and $E_{\Delta \cup \{D\}}^{\Lambda \cup \{D\}} = E_\Delta^\Lambda \cap D$. Therefore $E_\Delta^\Lambda = E_\Delta^{\Lambda \cup \{D\}} \cup E_{\Delta \cup \{D\}}^{\Lambda \cup \{D\}}$. $\qquad \square$

Lemma 3.5.23. *Let $A \in \mathfrak{B}(\Lambda)$, then $m_\Lambda^*(A) = m_{\Lambda \cup \{D\}}^*(A)$.*

Proof.

$$m_{\Lambda \cup \{D\}}^*(A) = \sum_{\{\Delta \subseteq \Lambda \cup \{D\} | E_\Delta^{\Lambda \cup \{D\}} \subseteq A\}} m_{\Lambda \cup \{D\}}^*(E_\Delta^{\Lambda \cup \{D\}})$$

$$= \sum_{\{\Delta \subseteq \Lambda | E_\Delta^{\Lambda \cup \{D\}} \subseteq A\}} m_{\Lambda \cup \{D\}}^*(E_\Delta^{\Lambda \cup \{D\}})$$

$$+ \sum_{\{\Delta \subseteq \Lambda | E_{\Delta \cup \{D\}}^{\Lambda \cup \{D\}} \subseteq A\}} m_{\Lambda \cup \{D\}}^*(E_{\Delta \cup \{D\}}^{\Lambda \cup \{D\}})$$

$$= \sum_{\{\Delta \subseteq \Lambda | E_\Delta^{\Lambda \cup \{D\}} \subseteq A\}} a_\Delta^{\Lambda \cup \{D\}} + \sum_{\{\Delta \subseteq \Lambda | E_{\Delta \cup \{D\}}^{\Lambda \cup \{D\}} \subseteq A\}} a_{\Delta \cup \{D\}}^{\Lambda \cup \{D\}}$$

$$= \sum_{\{\Delta \subseteq \Lambda | E_\Delta^\Lambda \subseteq A\}} (a_\Delta^{\Lambda \cup \{D\}} + a_{\Delta \cup \{D\}}^{\Lambda \cup \{D\}})$$

$$= \sum_{\{\Delta \subseteq \Lambda | E_\Delta^\Lambda \subseteq A\}} a_\Delta^\Lambda \text{ (Lemma 3.5.19)}$$

$$= m_\Lambda^*(A) \qquad \square$$

Corollary 3.5.24. *There is a unique finitely additive probability function m^* : $\mathfrak{B}(\Xi) \to \mathbb{R}$ such that for every finite $\Lambda \subseteq \Xi$ and $A \in \mathfrak{B}(\Xi)$, if $A \in \mathfrak{B}(\Lambda)$ then $m^*(A) = m^*_\Lambda(A)$.*

Lemma 3.5.25. *Such a m^* extends m.*

Proof. Suppose $A \in \Xi$. Then $A \in \mathfrak{B}(\{A\})$, and

$$m^*_{\{A\}}(A) = \sum_{E^\Lambda_\Delta \subseteq A} a^{\{A\}}_\Delta$$
$$= a^{\{A\}}_{\{A\}}$$
$$= m(A) - m(\varnothing) = m(A) \qquad \square$$

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \square$

To be able to use this theorem we will present a lemma telling us that we can pick a monotone 2-valuation to then apply this theorem to get a m_w. To show we can pick some such monotone 2-valuation we first need a pre-lemma that will generally be useful throughout the proof.

Lemma 3.5.26 (Goldblatt, 2014, Rich Extension III). *If Γ is $\vdash^\omega_{\mathsf{ProbKF}\cup\Sigma}$-consistent then there is some $w \in W^\Sigma_c$ such that $\Gamma \subseteq w$.*
Therefore if for every $w \in W^\Sigma_c$ $\varphi \in w$, then $\vdash^\omega_{\mathsf{ProbKF}\cup\Sigma} \varphi$.

Proof. To prove this we use the Henkin method. Enumerate the formulas with one free variable $\varphi_1(x), \varphi_2(x), \ldots$. We work by induction to find $\{\Gamma_n\}$ with

- Each Γ_n is $\vdash^\omega_{\mathsf{ProbKF}\cup\Sigma}$-consistent
- $\Gamma = \Gamma_0 \subseteq \Gamma_1 \subseteq \Gamma_2 \subseteq \ldots$
- Γ_{n+1} decides the ω-rule for φ_n, i.e.

  Either there is some k such that $\neg\varphi_n(\overline{k}) \in \Gamma_{n+1}$ or $\forall x \varphi_n(x) \in \Gamma_{n+1}$.

Suppose we have chosen Γ_n. Then suppose there is some k such that $\Gamma_n \cup \{\neg\varphi_n(\overline{k})\}$ is $\vdash^\omega_{\mathsf{ProbKF}\cup\Sigma}$-consistent. Then let $\Gamma_{n+1} = \Gamma_n \cup \{\neg\varphi_n(\overline{k})\}$ for this k. This clearly satisfies the requirements. If there is no such k, then for each k, $\Gamma_n \cup \{\neg\varphi_n(\overline{k})\}$ is $\vdash^\omega_{\mathsf{ProbKF}\cup\Sigma}$-inconsistent. So for each k, $\Gamma_n \vdash^\omega_{\mathsf{ProbKF}\cup\Sigma} \varphi_n(\overline{k})$, therefore $\Gamma_n \vdash^\omega_{\mathsf{ProbKF}\cup\Sigma} \forall x \varphi_n(x)$ using the ω-rule. Therefore we can let $\Gamma_{n+1} = \Gamma_n \cup \{\forall x \varphi_n(x)\}$, which is $\vdash^\omega_{\mathsf{ProbKF}\cup\Sigma}$-consistent because we assumed that Γ_n was.

We can then use Lindenbaum's lemma to show that $\bigcup \Gamma_n$ has a maximally finitely $\vdash^\omega_{\mathsf{ProbKF}\cup\Sigma}$-consistent extension w. This w will be closed under the ω-rule by construction.

For the second part of the theorem: Suppose for each $w \in W^\Sigma_c$, $\varphi \in w$, so $\neg\varphi \notin w$. Therefore by the contrapositive of the first result $\{\neg\varphi\}$ is $\vdash^\omega_{\mathsf{ProbKF}}$-inconsistent. So $\vdash^\omega_{\mathsf{ProbKF}} \varphi$. $\qquad\square$

We can now show that we can find a monotone 2-valuation that will be extended to be the m_w.

Lemma 3.5.27. *Fix $w \in W_c^\Sigma$, define*

$$[\varphi] := \{v \in W_c^\Sigma \mid \mathsf{T}^\ulcorner \varphi^\urcorner \in v\}$$

for $n \in \mathsf{Sent}_{\mathsf{P}_\geqslant, \mathsf{T}}$. Then $\{[\varphi] \mid \varphi \in \mathsf{Sent}_{\mathsf{P}_\geqslant, \mathsf{T}}\} \subseteq \wp(W_c^\Sigma)$ is a distributive lattice and that $m : \{[\varphi] \mid \varphi \in \mathsf{Sent}_{\mathsf{P}_\geqslant, \mathsf{T}}\} \to \mathbb{R}$ by $m([\varphi]) = \sup\{r \mid \mathsf{P}_\geqslant(^\ulcorner \varphi^\urcorner, ^\ulcorner r^\urcorner) \in w\}$ is a monotone 2-valuation on this.

Proof. We first show that this is a distributive lattice:

- By Axiom 2 for all $v \in W_c^\Sigma$ $\mathsf{T}^\ulcorner 0 = 0^\urcorner \in v$, so $[0 = 0] = W_c^\Sigma$.

- By Axiom 3 for all $v \in W_c^\Sigma$ $\neg\mathsf{T}^\ulcorner \neg 0 = 0^\urcorner \in v$, so $\mathsf{T}^\ulcorner \neg 0 = 0^\urcorner \notin v$, so $[\neg 0 = 0] = \varnothing$.

- Observe that $[\varphi] \cup [\psi] = [\varphi \vee \psi]$ because by Axiom 7 for all $v \in W_c^\Sigma$, $\mathsf{T}^\ulcorner \varphi \vee \psi^\urcorner \in v$ iff $\mathsf{T}^\ulcorner \varphi^\urcorner \in v$ or $\mathsf{T}^\ulcorner \psi^\urcorner \in v$.

- We can also show that $[\varphi] \cap [\psi] = [\varphi \wedge \psi]$ by showing that for all $v \in W_c^\Sigma$, $\mathsf{T}^\ulcorner \varphi \wedge \psi^\urcorner \in v$ iff both $\mathsf{T}^\ulcorner \varphi^\urcorner \in v$ and $\mathsf{T}^\ulcorner \psi^\urcorner \in v$. To show this it suffices to show that $\mathsf{T}^\ulcorner \varphi \wedge \psi^\urcorner \leftrightarrow \mathsf{T}^\ulcorner \varphi^\urcorner \wedge \mathsf{T}^\ulcorner \psi^\urcorner$ is derivable from the axioms of KF, or equivalently $\mathsf{T}^\ulcorner \neg(\neg\varphi \vee \neg\psi)^\urcorner \leftrightarrow (\mathsf{T}^\ulcorner \varphi^\urcorner \wedge \mathsf{T}^\ulcorner \psi^\urcorner)$. This works by using Axioms 6 and 8.

This has then shown that $\{[\varphi] \mid \varphi \in \mathsf{Sent}_{\mathsf{P}_\geqslant, \mathsf{T}}\} \subseteq \wp(W_c^\Sigma)$ is a distributive lattice. We now show that m is a monotone 2-valuation on it.

- To show $m(W_c^\Sigma) = 1$ it suffices to show that $\sup\{r \mid \mathsf{P}_\geqslant(^\ulcorner 0 = 0^\urcorner, ^\ulcorner r^\urcorner) \in w\} = 1$. By Axiom 19 $\mathsf{P}_\geqslant(^\ulcorner 0 = 0^\urcorner, ^\ulcorner 1^\urcorner) \in w$, so $m(W_c^\Sigma) \geqslant 1$. Also by Axiom 19, $\neg\mathsf{P}_>(^\ulcorner 0 = 0^\urcorner, ^\ulcorner 1^\urcorner) \in w$, i.e. $\neg\exists a \succ ^\ulcorner 1^\urcorner \mathsf{P}_\geqslant(^\ulcorner 0 = 0^\urcorner, a) \in w$. So for each $r > 1$, $^\ulcorner r^\urcorner \succ ^\ulcorner 1^\urcorner \in w$, so $\neg\mathsf{P}_\geqslant(^\ulcorner 0 = 0^\urcorner, ^\ulcorner r^\urcorner) \in w$.[35] Therefore $m(W_c^\Sigma) = 1$.

- The argument to show $m(\varnothing) = 0$ is directly analogous using Axiom 20.

- We now need to show: $[\varphi] \subseteq [\psi] \implies m[\varphi] \leqslant m[\psi]$. Suppose $[\varphi] \subseteq [\psi]$. Then for each $v \in W_c^\Sigma$, $\mathsf{T}^\ulcorner \varphi^\urcorner \to \mathsf{T}^\ulcorner \psi^\urcorner \in v$. Then by Lemma 3.5.26, $\vdash^\omega_{\mathsf{ProbKF}} \mathsf{T}^\ulcorner \varphi^\urcorner \to \mathsf{T}^\ulcorner \psi^\urcorner$, so using Rule 22 we have that $\vdash^\omega_{\mathsf{ProbKF}} \forall a(\mathsf{P}_\geqslant(^\ulcorner \varphi^\urcorner, a) \to \mathsf{P}_\geqslant(^\ulcorner \psi^\urcorner, b))$. So for any r, $\mathsf{P}_\geqslant(^\ulcorner \varphi^\urcorner, ^\ulcorner r^\urcorner) \in w \implies \mathsf{P}_\geqslant(^\ulcorner \psi^\urcorner, ^\ulcorner r^\urcorner) \in w$ i.e. $m[\varphi] \leqslant m[\psi]$.

- Lastly we need to show: $m([\varphi] \cup [\psi]) + m([\varphi] \cap [\psi]) = m([\varphi]) + m([\psi])$. $\forall b\forall c(\mathsf{P}_\geqslant(^\ulcorner \varphi^\urcorner, b) \wedge \mathsf{P}_\geqslant(^\ulcorner \psi^\urcorner, c) \to b \dotplus c \preceq ^\ulcorner r^\urcorner) \in w$ iff $r \geqslant m[\varphi] + m[\psi]$. And similarly $\forall d\forall e(\mathsf{P}_\geqslant(^\ulcorner \varphi \wedge \psi^\urcorner, d) \wedge \mathsf{P}_\geqslant(^\ulcorner \varphi \vee \psi^\urcorner, e) \to d \dotplus e \preceq a) \in w$ iff $r \geqslant m[\varphi \vee \psi] + m[\varphi \wedge \psi]$. Axiom 21 gives us for all $r \in \mathbb{Q}$,

$$(\forall b\forall c(\quad \mathsf{P}_\geqslant(^\ulcorner \varphi^\urcorner, b) \quad \wedge \quad \mathsf{P}_\geqslant(^\ulcorner \psi^\urcorner, c) \quad \to b \dotplus c \preceq ^\ulcorner r^\urcorner))$$
$$\leftrightarrow (\forall d\forall e(\mathsf{P}_\geqslant(^\ulcorner \varphi \wedge \psi^\urcorner, d) \wedge \mathsf{P}_\geqslant(^\ulcorner \varphi \vee \psi^\urcorner, e) \to d \dotplus e \preceq ^\ulcorner r^\urcorner))$$

  Using these we get that

$$\{r \in \mathbb{Q} \mid r \geqslant m[\varphi] + m[\psi]\} = \{r \in \mathbb{Q} \mid r \geqslant m[\varphi \vee \psi] + m[\varphi \wedge \psi]\}$$

  So $m[\varphi] + m[\psi] = m[\varphi \vee \psi] + m[\varphi \vee \psi]$ as required. $\qquad\square$

---

[35]Because for each $r > 1$, $\neg\exists a \succ ^\ulcorner 1^\urcorner \mathsf{P}_\geqslant(^\ulcorner 0 = 0^\urcorner, a) \vdash \neg\mathsf{P}_\geqslant(^\ulcorner 0 = 0^\urcorner, ^\ulcorner r^\urcorner)$, so using Item 7.

Using these two results we have our desired m_w.

Theorem 3.5.28. *For each $w \in W_c^\Sigma$ we can find m_w such that*

$$m_w(\{v \in W_c^\Sigma \mid \mathsf{T}\overline{n} \in v\}) = \sup\{r \mid \mathsf{P}_\geqslant(\overline{n}, \ulcorner r \urcorner) \in w\}.$$

Proof. Fix $w \in W_c^\Sigma$. Using Theorem 3.5.12 and Lemma 3.5.27 we have that there is some m defined on $\mathfrak{B}\left(\{[\varphi] \mid \varphi \in \mathsf{Sent}_{\mathsf{P}_\geqslant, \mathsf{T}}\}\right)$ which is probabilistic and extends μ as defined in Lemma 3.5.27. Such an m can then be extended to m_w defined on $\wp(W_c^\Sigma)$ by Proposition 1.2.2.

We need to show this m_w satisfies the required property. If $n \in \mathsf{Sent}_{\mathsf{P}_\geqslant, \mathsf{T}}$, $m_w(\{v \in W_c^\Sigma \mid \mathsf{T}\overline{n} \in v\}) = \sup\{r \mid \mathsf{P}_\geqslant(\overline{n}, \ulcorner r \urcorner) \in w\}$ holds by definition of μ in Lemma 3.5.27. Suppose $n \notin \mathsf{Sent}_{\mathsf{P}_\geqslant, \mathsf{T}}$. Then by Axiom 13, $\{v \mid \mathsf{T}\overline{n} \in v\} = \varnothing$, so $m_w\{v \mid \mathsf{T}\overline{n} \in v\} = 0$. By Axiom 3, $\vdash^\omega_{\mathsf{ProbKF} \cup \Sigma} \neg\mathsf{T}\ulcorner\neg 0 = 0\urcorner$, so $\vdash^\omega_{\mathsf{ProbKF} \cup \Sigma} \mathsf{T}\ulcorner\neg 0 = 0\urcorner \to \mathsf{T}\overline{n}$. By Rule 22, $\vdash^\omega_{\mathsf{ProbKF} \cup \Sigma} \mathsf{P}_\geqslant(\ulcorner\neg 0 = 0\urcorner, \ulcorner 0 \urcorner) \to \mathsf{P}_\geqslant(\overline{n}, \ulcorner 0 \urcorner)$. By Axiom 20, $\vdash^\omega_{\mathsf{ProbKF} \cup \Sigma} \mathsf{P}_\geqslant(\ulcorner\neg 0 = 0\urcorner, \ulcorner 0 \urcorner)$, so in fact $\vdash^\omega_{\mathsf{ProbKF} \cup \Sigma} \mathsf{P}_\geqslant(\overline{n}, \ulcorner 0 \urcorner)$. Therefore $\sup\{r \mid \mathsf{P}_\geqslant(\overline{n}, \ulcorner r \urcorner) \in w\} \geqslant 0$. We just need to show that it is also $\leqslant 0$. Suppose $\sup\{r \mid \mathsf{P}_\geqslant(\overline{n}, \ulcorner r \urcorner) \in w\} > 0$. Then there is some $q > 0$ with $\mathsf{P}_\geqslant(\overline{n}, \ulcorner q \urcorner) \in w$. This contradicts Axiom 17, so we get the result we need. $\square$

We have now shown that $\mathfrak{M}_c^\Sigma$ is a well-defined probabilistic modal structure. We will now show that it is canonical and that f is a fixed point.

$\mathfrak{M}_c^\Sigma$ is canonical and f is a fixed point

We can observe that this is canonical, in the following sense.

Lemma 3.5.29. *For every $\varphi \in \mathsf{Sent}_{\mathsf{P}_\geqslant, \mathsf{T}}$ and $w \in W_c^\Sigma$,*

$$\mathsf{IM}_{\mathfrak{M}_c^\Sigma}[w, f_c^\Sigma] \models \varphi \iff \varphi \in w$$

Proof. We work by induction on the complexity of the formula. The atomic cases with $\varphi \in \mathsf{Sent}_{\mathcal{L}}$ follow immediately from the definition of the canonical model. For the induction step the universal quantifier can be shown by the fact that $\mathsf{IM}_{\mathfrak{M}}[w, f]$ is an $\mathbb{N}$-model and w is closed under the ω-rule. For the connectives we use the fact that w is maximally consistent. For the $\mathsf{T}t$ case we use the definitions and Axiom 13. The most complex case is the $\mathsf{P}_\geqslant(t, s)$ case. If $s^\mathbb{N} \in \mathsf{Rat}$ and $t^\mathbb{N} \in \mathsf{Sent}_{\mathsf{P}_\geqslant, \mathsf{T}}$ then we can just use the definitions and Axioms 16, 18 and 18. If $s^\mathbb{N} \notin \mathsf{Rat}$ we also need to use Axiom 16. For if $t^\mathbb{N} \notin \mathsf{Sent}_{\mathsf{P}_\geqslant, \mathsf{T}}$ we also need to show: $\sup\{r \mid \mathsf{P}_\geqslant(\overline{t^\mathbb{N}}, \ulcorner r \urcorner) \in w\} = 0$, which we show by using Axioms 3, 17 and 20 and Rule 22, and $m_w\{v \mid \mathsf{T}\overline{n} \in v\} = 0$, which we do by using Axiom 13.

More carefully:

- φ atomic and in $\mathsf{Sent}_{\mathcal{L}}$:

$$\begin{aligned} \mathsf{IM}_{\mathfrak{M}_c^\Sigma}[w, f_c^\Sigma] \models \varphi &\iff \mathbf{M}(w) \models \varphi && \text{Definitions 3.2.3, 3.5.9 and 3.2.12} \\ &\iff \varphi \in w && \text{Definition 3.5.9} \end{aligned}$$

- $\varphi = \mathsf{T}t$:

$$
\begin{aligned}
\mathsf{IM}_{\mathfrak{M}_c^\Sigma}[w, f_c^\Sigma] \models \mathsf{T}t &\iff t^{\mathbb{N}} \in f_c^\Sigma(w) \text{ and } t^{\mathbb{N}} \in \mathsf{Sent}_{\mathsf{P}_\geqslant, \mathsf{T}} \\
&\iff \mathsf{T}\overline{t^{\mathbb{N}}} \in w \text{ and } t^{\mathbb{N}} \in \mathsf{Sent}_{\mathsf{P}_\geqslant, \mathsf{T}} && (3.7) \\
&\iff \mathsf{T}\overline{t^{\mathbb{N}}} \in w && (3.8) \\
&\iff \mathsf{T}t \in w
\end{aligned}
$$

For Eq. (3.7) $\implies$ Eq. (3.8): $\mathsf{T}\overline{t^{\mathbb{N}}} \in w \implies \mathsf{Sent}(\overline{t^{\mathbb{N}}}) \in w$ by Axiom 13. The result then follows because $\mathsf{Sent}_{\mathsf{P}_\geqslant, \mathsf{T}}$ strongly represents the set in PA.

- $\varphi = \mathsf{P}_\geqslant(t, s)$:

$$
\begin{aligned}
\mathsf{IM}_{\mathfrak{M}_c^\Sigma}[w, f_c^\Sigma] &\models \mathsf{P}_\geqslant(t, s) \\
&\iff m_w\{v \mid t^{\mathbb{N}} \in f_c^\Sigma(v)\} \geqslant \mathsf{rat}(s^{\mathbb{N}}) \text{ and } s^{\mathbb{N}} \in \mathsf{Rat} \\
&\iff m_w\{v \mid \mathsf{T}\overline{t^{\mathbb{N}}} \in v\} \geqslant \mathsf{rat}(s^{\mathbb{N}}) \text{ and } s^{\mathbb{N}} \in \mathsf{Rat} && (3.9) \\
&\iff \sup\{r \mid \mathsf{P}_\geqslant(\overline{t^{\mathbb{N}}}, \ulcorner r \urcorner) \in w\} \geqslant \mathsf{rat}(s^{\mathbb{N}}) \text{ and } s^{\mathbb{N}} \in \mathsf{Rat} && (3.10) \\
&\iff \mathsf{P}_\geqslant(\overline{t^{\mathbb{N}}}, \ulcorner \mathsf{rat}(s^{\mathbb{N}}) \urcorner) \in w \text{ and } s^{\mathbb{N}} \in \mathsf{Rat} && (3.11) \\
&\iff \mathsf{P}_\geqslant(t, s) \in w && (3.12)
\end{aligned}
$$

For Eq. (3.10) $\implies$ Eq. (3.11): Suppose $\sup\{r \mid \mathsf{P}_\geqslant(\overline{t^{\mathbb{N}}}, \ulcorner r \urcorner) \in w\} \geqslant \mathsf{rat}(s^{\mathbb{N}})$ and $s^{\mathbb{N}} \in \mathsf{Rat}$. Then by the definition of supremum, for each $q < \mathsf{rat}(s^{\mathbb{N}})$ there is an r with $q < r < \mathsf{rat}(s^{\mathbb{N}})$ with $\mathsf{P}_\geqslant(\overline{t^{\mathbb{N}}}, \ulcorner r \urcorner) \in w$. By using this and Axiom 18 we have that for all $r < \mathsf{rat}(s^{\mathbb{N}})$, $\mathsf{P}_\geqslant(\overline{t^{\mathbb{N}}}, \ulcorner r \urcorner) \in w$. So for all n, $(\overline{n} \prec \ulcorner \mathsf{rat}(s^{\mathbb{N}}) \urcorner \to \mathsf{P}_\geqslant(\overline{t^{\mathbb{N}}}, \overline{n})) \in w$. So since w is closed under the ω-rule, $\forall x(x \prec \ulcorner \mathsf{rat}(s^{\mathbb{N}}) \urcorner \to \mathsf{P}_\geqslant(\overline{t^{\mathbb{N}}}, x)) \in w$. Therefore by Axiom 18, $\mathsf{P}_\geqslant(\overline{t^{\mathbb{N}}}, \ulcorner \mathsf{rat}(s^{\mathbb{N}}) \urcorner) \in w$

For Eq. (3.12) $\implies$ Eq. (3.11): Suppose $\mathsf{P}_\geqslant(t, s) \in w$. Then by Axiom 16 $\mathsf{Rat}(s) \in w$. Since Rat strongly represents the corresponding set, $s^{\mathbb{N}} \in \mathsf{Rat}$. The equivalence Eq. (3.9) $\iff$ Eq. (3.10) holds by Definition 3.5.9.

We then use the induction hypothesis to show:

- $\varphi = \neg\psi$:

$$
\begin{aligned}
\mathsf{IM}_{\mathfrak{M}_c^\Sigma}[w, f_c^\Sigma] \models \neg\psi &\iff \mathsf{IM}_{\mathfrak{M}_c^\Sigma}[w, f_c^\Sigma] \not\models \psi \\
&\iff \psi \notin w && \text{Induction hypothesis} \\
&\iff \neg\psi \in w
\end{aligned}
$$

- $\varphi = \psi \wedge \chi$:

$$
\begin{aligned}
\mathsf{IM}_{\mathfrak{M}_c^\Sigma}[w, f_c^\Sigma] \models \psi \wedge \chi &\iff \mathsf{IM}_{\mathfrak{M}_c^\Sigma}[w, f_c^\Sigma] \models \psi \text{ and } \mathsf{IM}_{\mathfrak{M}_c^\Sigma}[w, f_c^\Sigma] \models \chi \\
&\iff \psi \in w \text{ and } \chi \in w \\
&\iff \psi \wedge \chi \in w
\end{aligned}
$$

- $\varphi = \forall x \psi(x)$:

$$\mathsf{IM}_{\mathfrak{M}_{\mathrm{c}}^{\Sigma}}[w, f_{\mathrm{c}}^{\Sigma}] \models \forall x \psi(x) \iff \mathsf{IM}_{\mathfrak{M}_{\mathrm{c}}^{\Sigma}}[w, f_{\mathrm{c}}^{\Sigma}] \models \psi(\overline{n}) \text{ for all } n \in \mathbb{N}$$
$$\iff \psi(\overline{n}) \in w \text{ for all } n \in \mathbb{N}$$
$$\iff \forall x \psi(x) \in w$$

The last equivalence holds because w is closed under the ω-rule. $\qquad \square$

The following lemma, along with Lemma 3.5.29 allows us to conclude that f_{c}^{Σ} is a fixed point, and therefore that the constructed model is in fact a probabilistic structure with a fixed point. This lemma is a slight modification of Stern (2014b, Theorem 3.13).

Lemma 3.5.30. *Let $\mathfrak{M}$ be a probabilistic modal structure and f be an evaluation function on $\mathfrak{M}$. Then:*

$$f \text{ is a fixed point} \iff \forall w \in W(\mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{KF} \cup \mathsf{InteractionAx})$$

Proof. The $\implies$-direction follows from Theorem 3.5.2.

For the other direction, suppose $\mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{KF} \cup \mathsf{InteractionAx}$.

Suppose $n \in f(w)$ so $\mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{T}\overline{n}$ by Definitions 3.2.3 and 3.2.12. By Axiom 17, we have $\mathsf{IM}_{\mathfrak{M}}[w, f] \models \forall x(\mathsf{T}x \to \mathsf{Sent}_{\mathsf{P}_{\geqslant}, \mathsf{T}}(x))$, so $n \in \mathsf{Sent}_{\mathsf{P}_{\geqslant}, \mathsf{T}}$.

This shows that $f(w) \subseteq \mathsf{Sent}_{\mathsf{P}_{\geqslant}, \mathsf{T}}$. We will work by induction on the positive complexity of φ to show that:

$$\mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{T}\ulcorner\varphi\urcorner \iff (w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \varphi$$

By Lemma 3.5.7 this will suffice to conclude that f is a fixed point.

We show this by induction on the positive complexity of φ:

- $\varphi \in \mathcal{L}$ atomic, i.e. φ is $s = t$ or $\varphi = Q(t_1, \ldots, t_n)$:
  $\quad \mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{T}\ulcorner\varphi\urcorner$
  iff $\quad \mathsf{IM}_{\mathfrak{M}}[w, f] \models \varphi \qquad$ Axioms 2 and 4,
  iff $\quad \mathbf{M}(w) \models \varphi \qquad$ Definition 3.2.12
  iff $\quad (w, f) \models_{\mathfrak{M}}^{SKP} \varphi \qquad$ Definition 3.2.3
- $\varphi = \neg\psi$ with $\psi \in \mathcal{L}$ atomic:
  $\quad \mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{T}\ulcorner\neg\psi\urcorner$
  iff $\quad \mathsf{IM}_{\mathfrak{M}}[w, f] \models \neg\psi \qquad$ Axioms 3 and 5
  iff $\quad \mathbf{M}(w) \models \neg\psi \qquad$ Definition 3.2.12
  iff $\quad (w, f) \models_{\mathfrak{M}}^{SKP} \neg\psi \qquad$ Definition 3.2.3
- $\varphi = \mathsf{P}_{\geqslant}(t, s)$
  $\quad \mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{T}\ulcorner\mathsf{P}_{\geqslant}(t, s)\urcorner$
  iff $\quad \mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{P}_{\geqslant}(t, s) \qquad$ Axiom 14
  iff $\quad (w, f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_{\geqslant}(t, s) \qquad$ Definition 3.2.12

- $\varphi = \neg\mathsf{P}_{\geqslant}(t,s)$

$\qquad \mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}^{\ulcorner}\neg\mathsf{P}_{\geqslant}(t,s)^{\urcorner}$

iff $\quad \mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{P}_{>}(\dot{\neg}t, 1\dot{-}s) \vee \neg\mathsf{Rat}(s)$ $\qquad$ Axiom 15

iff $\quad \mathsf{IM}_{\mathfrak{M}}[w,f] \models \exists a \succ 1\dot{-}s(\mathsf{P}_{\geqslant}(\dot{\neg}t, a)) \vee \neg\mathsf{Rat}(s)$

iff $\quad$ (there is some $r > \mathsf{rat}((1\dot{-}s)^{\mathbb{N}})$ with $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{P}_{\geqslant}(\dot{\neg}t, {}^{\ulcorner}r^{\urcorner}))$
$\qquad$ or $s^{\mathbb{N}} \notin \mathsf{Rat}$

iff $\quad$ (there is some $r > \mathsf{rat}((1\dot{-}s)^{\mathbb{N}})$ with $(w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{P}_{\geqslant}(\dot{\neg}t, {}^{\ulcorner}r^{\urcorner}))$
$\qquad$ or $s^{\mathbb{N}} \notin \mathsf{Rat}$

iff $\quad$ (there is some $r > \mathsf{rat}((1\dot{-}s)^{\mathbb{N}})$ with $m_w\{v \mid \dot{\neg}t^{\mathbb{N}} \in f(v)\} \geqslant r)$
$\qquad$ or $s^{\mathbb{N}} \notin \mathsf{Rat}$

iff $\quad m_w\{v \mid \dot{\neg}t^{\mathbb{N}} \in f(v)\} > 1 - \mathsf{rat}(s^{\mathbb{N}})$ or $s^{\mathbb{N}} \notin \mathsf{Rat}$
$\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\mathbb{Q}$ is dense in $\mathbb{R}$

iff $\quad (w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\mathsf{P}_{\geqslant}(t,s)$

- $\varphi = \mathsf{T}t$

$\qquad \mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}^{\ulcorner}\mathsf{T}t^{\urcorner}$

iff $\quad \mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}^{\ulcorner}t^{\urcorner\circ}$ $\qquad$ Axiom 11

iff $\quad \mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}t$ $\qquad$ definition of $^{\circ}$

iff $\quad (w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{T}t$ $\qquad$ Definition 3.2.12

- $\varphi = \neg\mathsf{T}t$

$\qquad \mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}^{\ulcorner}\neg\mathsf{T}t^{\urcorner}$

iff $\quad \mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}\dot{\neg}^{\ulcorner}t^{\urcorner\circ} \vee \neg\mathsf{Sent}_{\mathsf{P}_{\geqslant},\mathsf{T}}({}^{\ulcorner}t^{\urcorner\circ})$ $\qquad$ Axiom 12

iff $\quad (w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \mathsf{T}\dot{\neg}t$ or $t^{\mathbb{N}} \notin \mathsf{Sent}_{\mathsf{P}_{\geqslant},\mathsf{T}}$ $\qquad$ Definition 3.2.12

iff $\quad \dot{\neg}t^{\mathbb{N}} \in f(w)$ or $t^{\mathbb{N}} \notin \mathsf{Sent}_{\mathsf{P}_{\geqslant},\mathsf{T}}$ $\qquad$ Definition 3.2.3

iff $\quad (w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\mathsf{T}t$ $\qquad$ Definition 3.2.3

- $\varphi = \neg\neg\psi$

$\qquad \mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}^{\ulcorner}\neg\neg\psi^{\urcorner}$

iff $\quad \mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}^{\ulcorner}\psi^{\urcorner}$ $\qquad$ Axiom 6

iff $\quad (w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \psi$ $\qquad$ induction hypothesis

- $\varphi = \psi \vee \chi$

$\qquad \mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}^{\ulcorner}\psi \vee \chi^{\urcorner}$

iff $\quad \mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}^{\ulcorner}\psi^{\urcorner}$ or $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}^{\ulcorner}\chi^{\urcorner}$ $\qquad$ Axiom 7

iff $\quad (w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \psi$ or $(w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \chi$ $\qquad$ induction hypothesis

iff $\quad (w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \psi \vee \chi$ $\qquad$ Definition 3.2.3

- $\varphi = \neg(\psi \vee \chi)$

$\qquad \mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}^{\ulcorner}\neg(\psi \vee \chi)^{\urcorner}$

iff $\quad \mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}^{\ulcorner}\neg\psi^{\urcorner}$ and $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}^{\ulcorner}\neg\chi^{\urcorner}$ $\qquad$ Axiom 8

iff $\quad (w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\psi$ and $(w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\chi$ $\qquad$ induction hypothesis

iff $\quad (w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg(\psi \vee \chi)$ $\qquad$ Definition 3.2.3

- $\varphi = \exists v\psi$

$\qquad \mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}^{\ulcorner}\exists v\psi^{\urcorner}$

iff $\quad \mathsf{IM}_{\mathfrak{M}}[w,f] \models \exists y\mathsf{T}^{\ulcorner}\psi^{\urcorner}(y/v)$ $\qquad$ Axiom 9

iff $\quad$ there is some n with $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}^{\ulcorner}\psi[\overline{n}/v]^{\urcorner}$ $\qquad$ $\mathsf{IM}_{\mathfrak{M}}[w,f]$ is a standard model
$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ by Definition 3.2.12

iff $\quad$ there is some n with $(w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \psi[\overline{n}/v]$ $\qquad$ Induction Hypothesis

iff $\quad (w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \exists v\psi$ $\qquad$ Definition 3.2.3

- $\varphi = \neg\exists v\psi$

$$\mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}^{\ulcorner}\neg\exists v\psi^{\urcorner}$$

iff	$\mathsf{IM}_{\mathfrak{M}}[w,f] \models \forall y\mathsf{T}^{\ulcorner}\neg\psi^{\urcorner}(y/v)$	Axiom 10
iff	for all n, $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \mathsf{T}^{\ulcorner}\neg\psi[\overline{n}/v]^{\urcorner}$	$\mathsf{IM}_{\mathfrak{M}}[w,f]$ is a standard model
		by Definition 3.2.12
iff	for all n, $(w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\psi[\overline{n}/v]$	induction hypothesis
iff	$(w,f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\exists v\psi$	Definition 3.2.3 $\qquad\square$

Corollary 3.5.31. *f^{Σ}_{c} is a fixed point.*

Proof. For each $\varphi \in \mathsf{KF} \cup \mathsf{InteractionAx}$, $\vdash^{\omega}_{\mathsf{ProbKF}} \varphi$, and each $w \in W^{\Sigma}_{\mathsf{c}}$ is maximally $\vdash^{\omega}_{\mathsf{ProbKF}}$-consistent, so for each w, $\mathsf{KF} \cup \mathsf{InteractionAx} \subseteq w$. Then by Lemma 3.5.29, $\mathsf{IM}_{\mathfrak{M}^{\Sigma}_{\mathsf{c}}}[w,f^{\Sigma}_{\mathsf{c}}] \models \mathsf{KF} \cup \mathsf{InteractionAx}$. So by Lemma 3.5.30, f^{Σ}_{c} is a fixed point, as required. $\qquad\square$

The completeness theorem

This has now led us to our desired soundness and completeness theorems. There are a number of different forms that this can be written in.

Theorem 3.5.32. *The following are equivalent:*

1. *Δ is $\vdash^{\omega}_{\mathsf{ProbKF}\cup\Sigma}$-consistent*

2. *There is some $w \in W^{\Sigma}_{\mathsf{c}}$, such that $\mathsf{IM}_{\mathfrak{M}^{\Sigma}_{\mathsf{c}}}[w,f^{\Sigma}_{\mathsf{c}}] \models \Delta$.*

3. *There is some probabilistic modal structure $\mathfrak{M}$ with fixed point f, such that $\mathsf{IM}_{\mathfrak{M}}[f] \models \Sigma$, and for some $w \in W$, $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \Delta$*

Proof. Item 1 $\Longrightarrow$ Item 2:

Suppose Δ is $\vdash^{\omega}_{\mathsf{ProbKF}}$-consistent. Then by Lemma 3.5.26 there is some $w \in W^{\Sigma}_{\mathsf{c}}$ such that $\Delta \subseteq w$. Then by Lemma 3.5.29 $\mathsf{IM}_{\mathfrak{M}^{\Sigma}_{\mathsf{c}}}[w,f^{\Sigma}_{\mathsf{c}}] \models \Delta$. Moreover we have shown in Corollary 3.5.31 that f^{Σ}_{c} is a fixed point. So we have our relevant probabilistic modal structure, $\mathfrak{M}^{\Sigma}_{\mathsf{c}}$, fixed point f^{Σ}_{c} and $w \in W^{\Sigma}_{\mathsf{c}}$ which satisfies Δ.

Clearly Item 2 $\Longrightarrow$ Item 3.

For Item 3 $\Longrightarrow$ Item 1 we prove the contrapositive. Suppose Δ is not $\vdash^{\omega}_{\mathsf{ProbKF}\cup\Sigma}$-consistent, i.e. $\Delta \vdash^{\omega}_{\mathsf{ProbKF}\cup\Sigma} \bot$. Then by Theorem 3.5.8 for every $\mathfrak{M}$ probabilistic modal structure and fixed point f with $\mathsf{IM}_{\mathfrak{M}}[f] \models \Sigma$, it must be that $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \Delta \Longrightarrow \mathsf{IM}_{\mathfrak{M}}[w,f] \models \bot$. Since $\mathsf{IM}_{\mathfrak{M}}[w,f] \not\models \bot$ it must be that $\mathsf{IM}_{\mathfrak{M}}[w,f] \not\models \Delta$. Therefore there is no such $\mathfrak{M}, w, f$ with $\mathsf{IM}_{\mathfrak{M}}[w,f] \models \Delta$, as required. $\qquad\square$

Corollary 3.5.33 (Theorem 3.5.5)**.** *The following are equivalent:*

1. *$\Gamma \vdash^{\omega}_{\mathsf{ProbKF}\cup\Sigma} \varphi$,*

2. *For every $w \in W^{\Sigma}_{\mathsf{c}}$, $\mathsf{IM}_{\mathfrak{M}^{\Sigma}_{\mathsf{c}}}[w,f^{\Sigma}_{\mathsf{c}}] \models \Gamma \Longrightarrow \mathsf{IM}_{\mathfrak{M}^{\Sigma}_{\mathsf{c}}}[w,f^{\Sigma}_{\mathsf{c}}] \models \varphi$.*

3. *For every probabilistic modal structure $\mathfrak{M}$ with fixed point f, such that $\mathsf{IM}_{\mathfrak{M}}[f] \models \Sigma$, we have that for each $w \in W$,*

$$\mathsf{IM}_{\mathfrak{M}}[w,f] \models \Gamma \Longrightarrow \mathsf{IM}_{\mathfrak{M}}[w,f] \models \varphi.$$

Proof. By taking the negation of each of the equivalent clauses in Theorem 3.5.32 we have that the following are equivalent.

1. $\Gamma \cup \{\neg\varphi\}$ is $\vdash^\omega_{\mathsf{ProbKF}\cup\Sigma}$-inconsistent
2. There is no $w \in W^\Sigma_c$, such that $\mathsf{IM}_{\mathfrak{M}^\Sigma_c}[w, f^\Sigma_c] \models \Gamma \cup \{\neg\varphi\}$.
3. There is no probabilistic modal structure $\mathfrak{M}$ and fixed point f, such that $\mathsf{IM}_{\mathfrak{M}}[f] \models \Sigma$, and for some $w \in W$, $\mathsf{IM}_{\mathfrak{M}}[w, f] \models \Gamma \cup \{\neg\varphi\}$

The required equivalences directly follows from this. □

Theorem 3.5.34 (Theorem 3.5.6)**.** *Let* $\mathcal{M}$ *be a* $\mathcal{L}_{\mathsf{P}_\geqslant,\mathsf{T}}$*-model.*

1. *The following are equivalent:*

   (a) $\mathcal{M} \models \Gamma \implies \mathcal{M} \models \varphi$ *whenever* $\Gamma \vdash^\omega_{\mathsf{ProbKF}\cup\Sigma} \varphi$,

   (b) $\mathsf{Theory}(\mathcal{M})$ *is* $\vdash^\omega_{\mathsf{ProbKF}\cup\Sigma}$*-consistent,*

   (c) *There is an probabilistic structure* $\mathfrak{M}$*, fixed point* f *such that* $\mathsf{IM}_{\mathfrak{M}}[f] \models \Sigma$*, and there is some* $w \in W$ *such that* $\mathcal{M}$ *is elementarily equivalent to* $\mathsf{IM}_{\mathfrak{M}}[w, f]$.[36]

2. *Suppose* $\mathcal{M}$ *is an* $\mathbb{N}$*-model.*[37] *Then the following are equivalent:*

   (a) $\mathcal{M} \models \varphi$ *for each* $\vdash^\omega_{\mathsf{ProbKF}} \varphi$,

   (b) $\mathsf{Theory}(\mathcal{M})$ *is finitely* $\vdash^\omega_{\mathsf{ProbKF}\cup\Sigma}$*-consistent,*

   (c) *There is an probabilistic structure* $\mathfrak{M}$*, fixed point* f *such that* $\mathsf{IM}_{\mathfrak{M}}[f] \models \Sigma$*, and there is some* $w \in W$ *with* $\mathcal{M} = \mathsf{IM}_{\mathfrak{M}}[w, f]$.

Proof. Item 1b $\iff$ Item 1c directly follows from Theorem 3.5.32 taking $\Delta = \mathsf{Theory}(\mathcal{M})$ and observing that $\mathsf{IM}_{\mathfrak{M}}[w, f]$ is elementarily equivalent to $\mathcal{M}$ iff $\mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{Theory}(\mathcal{M})$.

For Item 1a $\implies$ Item 1b we can prove the contrapositive. Suppose $\mathsf{Theory}(\mathcal{M})$ is $\vdash^\omega_{\mathsf{ProbKF}\cup\Sigma}$-inconsistent. Then $\mathsf{Theory}(\mathcal{M}) \vdash^\omega_{\mathsf{ProbKF}\cup\Sigma} \bot$. Now $\mathcal{M} \models \mathsf{Theory}(\mathcal{M})$, and $\mathcal{M} \not\models \bot$, which suffices for the result.

For Item 1a $\impliedby$ Item 1b: Suppose $\mathsf{Theory}(\mathcal{M})$ is $\vdash^\omega_{\mathsf{ProbKF}\cup\Sigma}$-consistent, $\mathcal{M} \models \Gamma$ and $\mathcal{M} \not\models \varphi$. Then $\Gamma \cup \{\neg\varphi\}$ is $\vdash^\omega_{\mathsf{ProbKF}\cup\Sigma}$-consistent, so $\Gamma \not\vdash^\omega_{\mathsf{ProbKF}\cup\Sigma} \varphi$.

Item 2c $\implies$ Item 2b is a direct corollary of Item 1c $\implies$ Item 1b.

For Item 2b $\implies$ Item 2a: Suppose $\mathsf{Theory}(\mathcal{M})$ is finitely $\vdash^\omega_{\mathsf{ProbKF}\cup\Sigma}$-consistent and $\vdash^\omega_{\mathsf{ProbKF}\cup\Sigma} \varphi$. If $\mathcal{M} \not\models \varphi$, then $\neg\varphi \in \mathsf{Theory}(\mathcal{M})$ but since $\neg\varphi \vdash^\omega_{\mathsf{ProbKF}\cup\Sigma} \bot$, $\mathsf{Theory}(\mathcal{M})$ is finitely $\vdash^\omega_{\mathsf{ProbKF}\cup\Sigma}$-inconsistent. Therefore in fact $\mathcal{M} \models \varphi$.

For Item 2a $\implies$ Item 2c: Suppose Item 2a and that $\mathcal{M}$ is an $\mathbb{N}$-model. Then $\mathsf{Theory}(\mathcal{M})$ is finitely $\vdash^\omega_{\mathsf{ProbKF}\cup\Sigma}$-consistent (it is also maximally so because for every φ, $\varphi \in \mathsf{Theory}(\mathcal{M})$ or $\neg\varphi \in \mathsf{Theory}(\mathcal{M})$). It is also closed under the ω-rule because it is an $\mathbb{N}$-model. So $\mathsf{Theory}(\mathcal{M}) \in W^\Sigma_c$. Therefore by Lemma 3.5.29, $\mathsf{IM}_{\mathfrak{M}^\Sigma_c}[\mathsf{Theory}(\mathcal{M}), f^\Sigma_c] \models \mathsf{Theory}(\mathcal{M})$. So $\mathsf{IM}_{\mathfrak{M}^\Sigma_c}[\mathsf{Theory}(\mathcal{M}), f^\Sigma_c]$ and $\mathcal{M}$ are elementarily equivalent. Since they are both $\mathbb{N}$-models they must in fact be identical. □

---

[36]I.e. $\mathcal{M}$ and $\mathsf{IM}_{\mathfrak{M}}[w, f]$ satisfy all the same $\mathcal{L}_{\mathsf{P}_\geqslant,\mathsf{T}}$-sentences.

[37]We still need this assumption because by assumption all $\mathsf{IM}_{\mathfrak{M}}[w, f]$ are $\mathbb{N}$-models, but even adding the ω-rule does not fix the standard model of arithmetic, it only fixes the *theory* of the standard model of arithmetic.

3.5.3 Adding additional axioms – consistency and introspection

The axiom system KF is often stated with a consistency axiom which succeeds in picking out the *consistent* fixed points. We have stated KF without the consistency axiom, following Halbach (2014). However, we can easily add the consistency axiom back in by taking it as a member of Σ.

Definition 3.5.35. Let ProbKFC denote ProbKF with the additional axiom

$$\forall x(\mathsf{Sent}_{\mathsf{P}_{\geqslant},\mathsf{T}}(x) \to \neg(\mathsf{T}x \wedge \mathsf{T}_{\dot\neg}x)). \tag{Cons}$$

Lemma 3.5.36. *Then* $\mathsf{IM}_{\mathfrak{M}}[f] \models \forall x(\mathsf{Sent}_{\mathsf{P}_{\geqslant},\mathsf{T}}(x) \to \neg(\mathsf{T}x \wedge \mathsf{T}_{\dot\neg}x))$ *if and only if f is consistent.*

Theorem 3.5.37. *The following are equivalent:*[38]

- $\Gamma \vdash^{\omega}_{\mathsf{ProbKFC}} \varphi$,

- *For every probabilistic modal structure $\mathfrak{M}$ with consistent fixed point f, we have that for each $w \in W$,*

$$\mathsf{IM}_{\mathfrak{M}}[w, f] \models \Gamma \implies \mathsf{IM}_{\mathfrak{M}}[w, f] \models \varphi.$$

We might also consider introspective frames now by using an axiom of introspection expressed using the truth predicate Section 3.4.1.

Definition 3.5.38. Consider the axiom scheme (Intro):

$$\mathsf{T}^\ulcorner \mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner \to \mathsf{P}_{=}(\ulcorner\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner, \ulcorner 1\urcorner)$$
$$\wedge \quad \mathsf{T}^\ulcorner \neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner \to \mathsf{P}_{=}(\ulcorner\neg\mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner r\urcorner)\urcorner, \ulcorner 1\urcorner)$$

Using Proposition 3.4.2, we obtain the following theorem as as a special case of Theorem 3.5.6.

Theorem 3.5.39. *Suppose $\mathcal{L}$ contains at least one empirical symbol. Then the following are equivalent:*

- $\Gamma \vdash^{\omega}_{\mathsf{ProbKF} \cup \{\mathrm{Intro}\}} \varphi$,

- *For every weakly introspective probabilistic modal structure $\mathfrak{M}$ with fixed point f, we have that for each $w \in W$,*

$$\mathsf{IM}_{\mathfrak{M}}[w, f] \models \Gamma \implies \mathsf{IM}_{\mathfrak{M}}[w, f] \models \varphi.$$

---

[38]And so is: For every $w \in W_{\mathsf{c}}^{\{\mathrm{Cons}\}}$,

$$\mathsf{IM}_{\mathfrak{M}_{\mathsf{c}}^{\{\mathrm{Cons}\}}}[w, f_{\mathsf{c}}^{\{\mathrm{Cons}\}}] \models \Gamma \implies \mathsf{IM}_{\mathfrak{M}_{\mathsf{c}}^{\{\mathrm{Cons}\}}}[w, f_{\mathsf{c}}^{\{\mathrm{Cons}\}}] \models \varphi.$$

3.6   Conclusions

In this chapter we have presented a construction of a semantics for a language that includes sentences that can talk about their own probabilities and have given a corresponding axiomatic theory. The semantics is developed by applying a familiar construction of a semantics for type-free truth, namely Kripke's construction from Kripke (1975), to possible world style structures. In this semantics some sentences are only assigned ranges of probability values instead of a single value but this will only happen for "problematic" sentences. In most cases, sentences one wants to work with will be grounded so they will then be assigned a particular probability value and one can reason in a fairly natural way. We provided an axiomatisation that allows one to reason about these semantics in a clear way. One could also use this axiomatisation to show what assumptions about probability would lead to inconsistencies.

We showed that if one expresses introspection principles by using a truth predicate to do the job of quotation and disquotation these introspection principles are consistent. Although we have only considered introspection principles here, we believe the phenomenon is quite general. For evidence of this we can see in Stern (2014a,b) that the strategy worked well in the case of necessity. In future work we would like to investigate exactly how one should express principles in order to avoid the paradoxical contradictions.

This construction does not yet have the ability to account for conditional probabilities. Furthermore, it is not clear that it would be possible to add conditional probabilities and give a good definition of $(w,f) \models_{\mathfrak{M}}^{\mathsf{SKP}} \mathsf{P}_{\geqslant r}(\ulcorner\varphi\urcorner|\ulcorner\psi\urcorner)$ in the style of strong Kleene three valued scheme. However, in this thesis we are not considering conditional probabilities (see Section 1.7). One might overcome this limitation by instead using a supervaluational evaluation scheme, we turn to that option in a short next chapter.

Chapter 4

A Supervaluational Kripke Construction and Imprecise Probabilities

In the previous chapter we considered a Kripkean construction based on a strong Kleene style evaluation scheme, as defined in Definition 3.2.3. In this chapter we consider an alternative that is based on supervaluational logic. The particular reason that this version is interesting is that it ends up bearing a nice connection to *imprecise probabilities*. Imprecise probabilities is a model of probability which drops some particular assumptions of traditional probability, often by modelling probability by *sets of probability measures*. It is a model that has been suggested for many reasons, for example because numerically precise credences are psychologically unrealistic, imprecise evidence may best be responded to by having imprecise credences, and they can represent incomparability in an agent's beliefs in a way that precise probabilities cannot. For an introduction to imprecise probabilities Bradley (2015) can be consulted.

As we saw in Section 2.4, for many probabilistic modal structures there are no classical valuations. We might alternatively describe this as saying there are no *stable states*: whatever probability evaluation function is chosen, some other probability evaluation function looks better from the original function's perspective, i.e. $\Theta(\mathbf{p}) \neq \mathbf{p}$. It turns out that this is not the case in the imprecise case. There are some imprecise probability assignments which *do* look best from their own perspective, i.e. there are some stable states. This can therefore be seen as an argument for imprecise probabilities that is very different from the existing arguments.

Outline of the chapter

In Section 4.1.1 we will present the formal definition of the semantics, which is a Kripke-style construction using ideas from supervaluational logic and imprecise probabilities. Then in Section 4.1.2 we consider some examples which give the idea of how this semantics works.

In Section 4.2 we apply these considerations to a different problem: how to provide a semantics to a group of imprecise reasoners reasoning about one

another. Here we are working with the operator language which contains operators $\mathbb{P}^{A}_{\geqslant r}$. To do this we will generalise the notion of a probabilistic modal structure to an *imprecise* probabilistic modal structures. An interesting feature of the semantics is that if an agent reasons about an imprecise agent then the reasoner will also be imprecise.

In the final section, Section 4.3 we will discuss the issue of convexity. It is often argued that imprecise probabilities should work with *convex* sets of probabilities as these have a stronger behavioural motivation, but the semantics that we have worked with ends up with non-convex probability sets as stable. One could alter the semantics definition to obtain convex sets, but we will not suggest doing that because that would mean we loose the property that the resulting stable sets will have that every member of the credal state looks best from some (possibly distinct) member of the credal state's perspective. That is a feature of our semantics that we think is very important, so we do not advice altering the semantics to obtain convex sets.

4.1 The semantics and stable states

4.1.1 Developing the semantics

When we are working with supervaluational logic we can be more flexible with the language. Now, instead of considering $\mathcal{L}_{\mathsf{P}_{\geqslant},\mathsf{T}}$ as in Definition 1.6.10, which worked with a binary predicate and a coding of rational numbers, we can work with a language which has P as a function symbol and works with a background theory of reals and arithmetic.

Setup 4 (for Chapter 4). *Let $\mathcal{L}$ be some first-order language extending $L_{\mathsf{PA},\mathsf{ROCF}}$ with predicates N and R standing for the natural numbers and real numbers, respectively. We will consider $\mathcal{L}_{\mathsf{P}}$, which extends $\mathcal{L}$ by the addition of a function symbol P.*

See Definition 1.6.11 for the introduction to this language and a discussion of it.

We here develop a construction which can be seen as a supervaluational version of Kripke's theory. In supervaluational logic one considers a number of permissible interpretations of a language. In our case we have different worlds, so we want to consider collections of permissible interpretations at each world. We do this by considering sets of functions assigning an interpretation at each world.

In the previous chapter we considered an evaluation function, which works as the extension of T at each world, and used that to determine the extension of P. We could therefore use collections of evaluation functions to provide a collection of permissible interpretations of $\mathcal{L}_{\mathsf{P}}$. However, in this chapter we are not interested in coordinating P with T, we are just interested in the interpretation of P, so we can rephrase it directly with sets of probabilities.[1] This also allows our discussion to fit closer with the existing literature on imprecise probabilities.

We already introduced prob-eval functions in Definition 2.4.1. This provides an interpretation of P at each world. Formally **p** is some function assigning to

---

[1]This construction in terms of prob-evaluation functions is equivalent to that one would define using collections of usual evaluation functions. See Section 4.A.

each world a function from $\mathsf{Sent}_{\mathcal{L}_\mathsf{P}}$ to $\mathbb{R}$. If there are multiple agents, a prob-evaluation function would be a collection of prob-evaluation functions, one for each agent. And we would represent this as $\langle \mathbf{p}^A \rangle$.

We are interested in this chapter in the imprecise case, so we define imprecise evaluations to be given by collections of precise ones.

Definition 4.1.1. An *imprec-prob-eval function*, $\boldsymbol{\mathcal{P}}$, is a collection of prob-evaluation functions.

This provides a collection of permissible interpretations of P for each agent at each world and coordinates these interpretations.

We can now give an operation, Θ, where, if $\boldsymbol{\mathcal{P}}$ is the current imprec-prob-eval function, $\Theta(\boldsymbol{\mathcal{P}})$ gives the imprec-prob-eval function at the next stage. This will be the imprec-prob-eval function that looks best from $\boldsymbol{\mathcal{P}}$'s perspective.

Remember we defined $\Theta_{\mathfrak{M}}$ for $\mathbf{p}$ in Definition 2.4.4 by:

$$\Theta_{\mathfrak{M}}(\mathbf{p})(w)(\varphi) = m_w\{v \mid (\mathbf{M}, \mathbf{p})(w) \models \varphi\}.$$

We can then use this to directly determine Θ for imprec-prob-eval functions.

Definition 4.1.2. Define

$$\Theta_{\mathfrak{M}}(\boldsymbol{\mathcal{P}}) = \{\Theta_{\mathfrak{M}}(\mathbf{p}) \mid \mathbf{p} \in \boldsymbol{\mathcal{P}}\}.$$

As usual, we will generally drop the explicit reference to $\mathfrak{M}$.

Just as in the Kripkean version, Θ is monotone so it has fixed points.

Proposition 4.1.3. Θ *is monotone. I.e. if* $\boldsymbol{\mathcal{P}} \subseteq \boldsymbol{\mathcal{P}}'$ *then* $\Theta(\boldsymbol{\mathcal{P}}) \subseteq \Theta(\boldsymbol{\mathcal{P}}')$.

Proof. Suppose $\boldsymbol{\mathcal{P}} \subseteq \boldsymbol{\mathcal{P}}'$.

$$\begin{aligned}
\Theta(\boldsymbol{\mathcal{P}}) &= \{\Theta(\mathbf{p}) \mid \mathbf{p} \in \boldsymbol{\mathcal{P}}\} \\
&\subseteq \{\Theta(\mathbf{p}) \mid \mathbf{p} \in \boldsymbol{\mathcal{P}}'\} \\
&= \Theta(\boldsymbol{\mathcal{P}}') \qquad\qquad\qquad\qquad \square
\end{aligned}$$

Corollary 4.1.4. *For every* $\mathfrak{M}$, *there is some* $\boldsymbol{\mathcal{P}}$ *that is a fixed point of* $\Theta_{\mathfrak{M}}$.

Definition 4.1.5. If $\boldsymbol{\mathcal{P}}$ is a fixed point we call it a *stable state*.

We informally describe such stable states as the ones that look best from their own perspectives. This is because we suggest that $\Theta(f)$ is the prob-eval function that looks best from f's perspective. A suggestion for a justification of this informal way of speaking is considered in Section 7.3.4 where what it means to look best from one's perspective is to minimize estimated inaccuracy, where the estimation is taken according to the accessibility measure, but using the current prob-eval function to interpret P at the different worlds.

So imprecise probabilities, as sets of probability functions, can provide stable states, whereas single probability functions cannot.

We just need to check that the members of the stable states are in fact probabilistic.

Proposition 4.1.6. *If* $\boldsymbol{\mathcal{P}}$ *is a stable state then each* $\mathbf{p} \in \boldsymbol{\mathcal{P}}$ *is probabilistic.*

Proof. Let $\mathbf{p} \in \boldsymbol{\mathcal{P}}$ and $\boldsymbol{\mathcal{P}}$ be a stable state. Then there is some $\mathbf{p}' \in \boldsymbol{\mathcal{P}}$ with $\Theta(\mathbf{p}') = \mathbf{p}$, so

$$\mathbf{p}(w)(\varphi) = \Theta(\mathbf{p}')(w)(\varphi) = m_w\{v \mid (\mathbf{M}, \mathbf{p}')(v) \models \varphi\}.$$

So $\mathbf{p}(w)$ is given by a probabilistic measure over classical models, and $\mathbf{p}(w)$ must therefore be probabilistic (see Proposition 1.2.4). □

4.1.2 Examples

Let's take a closer look at how these work.

We will write

$$\boldsymbol{\mathcal{P}}(w)(\varphi) := \{\mathbf{p}(w)(\varphi) \mid \mathbf{p} \in \boldsymbol{\mathcal{P}}\},$$

though there is more information in $\boldsymbol{\mathcal{P}}$ than just the $\boldsymbol{\mathcal{P}}(w)(\varphi)$, for example it might only contain prob-eval functions where the probability of ψ at w plus the probability of φ at w equals 1, but $\boldsymbol{\mathcal{P}}(w)(\varphi) = \boldsymbol{\mathcal{P}}(w)(\psi) = [0, 1]$.

Consider the degrees of belief of an agent who is omniscient about non-semantic state of affairs and is introspective. She is then represented in the very simple trivial probabilistic modal structure, $\mathfrak{M}_{\mathsf{omn}}$:
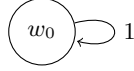
Example 4.1.7. Consider π with $\mathsf{PA} \vdash \pi \leftrightarrow \neg \mathsf{P}^\ulcorner \pi \urcorner \geqslant {}^1\!/_2$ in $\mathfrak{M}_{\mathsf{omn}}$.

Suppose $\mathbf{p}(w_0)(\pi) \not\geqslant {}^1\!/_2$. Then $(\mathbf{M}, \mathbf{p})(w_0) \models \pi$. So

$$\begin{aligned}
\Theta(\mathbf{p})(w_0)(\pi) &= m_{w_0}\{v \mid (\mathbf{M}, \mathbf{p})(v) \models \pi\} \\
&= m_{w_0}\{w_0\} \\
&= 1
\end{aligned}$$

Suppose $\mathbf{p}(w_0)(\pi) \geqslant {}^1\!/_2$. Then $(\mathbf{M}, \mathbf{p})(w_0) \models \neg\pi$. So

$$\Theta(\mathbf{p})(w_0)(\pi) = 0$$

So a stable state will have

$$\boldsymbol{\mathcal{P}}(w_0)(\pi) = \{0, 1\}$$

This is stable because the probability value of 1 looks best from 0's perspective, and 0 looks best from 1's perspective.

It is important that we consider this as that the whole set gives the appropriate interpretation of probability. It could not be that there some particular member of the set which is the correct interpretation but it is indeterminate which. This is because each member of the set looks bad from its own perspective, but there's a coherence about the whole set: every member is endorsed by some (often different) member.[2]

Sometimes there are multiple stable states:

---

[2]This does then lead to the following challenge that was pointed out to me by Jack Spencer: if the agents mental states really are the whole set of probabilities, then that is the sort of thing that an introspective agent should be able to recognise. This will then mean that the introspection conflict rearises. We would also then want to work with a language that can refer to imprecise credal states. Further analysis of this is left to future research.

Example 4.1.8. Suppose $Handstand \leftrightarrow \mathsf{P}\ulcorner Handstand \urcorner \geqslant 1/2$ is satisfied in $\mathfrak{M}_{\mathsf{omn}}$.[3] This formalises a setup where if the agent has confidence in her abilities she will be able to do the handstand, but if she doubts herself then she will not. This is closely related to the truthteller $\tau \leftrightarrow \mathsf{T}\ulcorner \tau \urcorner$.

Suppose $\mathbf{p}(w_0)(Handstand) \not\geqslant 1/2$. Then $(\mathbf{M}, \mathbf{p})(w_0) \models \neg\mathsf{P}\ulcorner Handstand \urcorner \geqslant 1/2$, so $(\mathbf{M}, \mathbf{p})(w_0) \models \neg Handstand$. So

$$\Theta(\mathbf{p})(w_0)(Handstand) = m_{w_0}\{v \mid (\mathbf{M}, \mathbf{p})(v) \models Handstand\}$$
$$= 0$$

Suppose $\mathbf{p}(w_0)(Handstand) \geqslant 1/2$. Then $(\mathbf{M}, \mathbf{p})(w_0) \models Handstand$. So

$$\Theta(\mathbf{p})(w_0)(Handstand) = 1$$

So, at least as far as *Handstand* goes, there are three stable states:

- $\boldsymbol{\mathcal{P}}(w_0)(Handstand) = 0$

- $\boldsymbol{\mathcal{P}}(w_0)(Handstand) = 1$

- $\boldsymbol{\mathcal{P}}(w_0)(Handstand) = \{0, 1\}$

We can also now consider empirical self-reference as discussed in Caie (2013).

Example 4.1.9. Suppose we have a probabilistic modal structure with some sentence *FreeThrow* as described in Fig. 4.1.
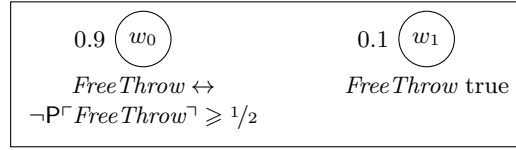
Figure 4.1: The $\mathfrak{M}$ in Example 4.1.9

Suppose $\mathbf{p}(w_0)(FreeThrow) \not\geqslant 1/2$. Then $(\mathbf{M}, \mathbf{p})(w_0) \models FreeThrow$. So

$$\Theta(\mathbf{p})(w_0)(FreeThrow) = m_{w_0}\{v \mid (\mathbf{M}, \mathbf{p})(v) \models FreeThrow\}$$
$$= m_{w_0}\{w_0, w_1\}$$
$$= 1$$

Suppose $\mathbf{p}(w_0)(FreeThrow) \geqslant 1/2$. Then $(\mathbf{M}, \mathbf{p})(w_0) \models \neg FreeThrow$. So

$$\Theta(\mathbf{p})(w_0)(FreeThrow) = m_{w_0}\{v \mid (\mathbf{M}, \mathbf{p})(v) \models FreeThrow\}$$
$$= m_{w_0}\{w_1\}$$
$$= 0.1$$

So, a stable evaluation will have: $\boldsymbol{\mathcal{P}}(w_0)(FreeThrow) = \{0.1, 1\}$.

---

[3]Or at least where it is presupposed that for any prob-eval function, $\mathbf{p}$, $(\mathbf{M}, \mathbf{p})(w_0) \models Handstand \iff (\mathbf{M}, \mathbf{p})(w_0) \models \mathsf{P}\ulcorner Handstand \urcorner \geqslant 1/2$. We might just work with the formal sentence η where $\mathsf{PA} \vdash \eta \leftrightarrow \mathsf{P}\ulcorner \eta \urcorner \geqslant 1/2$, which would be immune to Egan and Elga (2005)'s suggestion that equivalences between statements and the agent's belief in them should not be known by the agent and therefore not modelled as true in every world in the probabilistic modal structure.

We can also use these considerations and this semantics to analyse situations without self-reference but where instead there is a group of imprecise agents reasoning about one another's beliefs, and where the language is just the operator language $\mathcal{L}_{\mathbb{P}_{\geqslant r}}$.

4.2 Semantics for embedded imprecise probabilities

This semantics we have provided can be used and interesting, even in the operator case.

We will here talk about an agent's belief state being modelled by a *credal committee*, which consists of a set of probability functions.

Example 4.2.1. Consider the operator language $\mathcal{L}_{\mathbb{P}_{\geqslant r}^{\mathrm{Owen}}, \mathbb{P}_{\geqslant r}^{\mathrm{Ann}}}$ as set up in Definition 1.6.14, which will have sentences like:

$$\mathbb{P}^{\mathrm{Owen}}(\neg \mathbb{P}_{\geqslant 1/2}^{\mathrm{Ann}}(Heads)) = 1.$$

To determine the truth of this we need to determine:

$$\mathrm{p}^{\mathrm{Owen}}(\neg \mathbb{P}_{\geqslant 1/2}^{\mathrm{Ann}}(Heads)) = ?$$

Suppose Ann has $\mathrm{p}^{\mathrm{Ann}}(Heads) = \{0.1, 0.9\}$, and Owen knows this.

When Owen considers the member of Ann's credal committe that has degree of belief 0 in *Heads*, he should have degree of belief 1 in $\neg \mathbb{P}_{\geqslant 1/2}^{\mathrm{Ann}}(Heads)$. When he considers the other member of her credal committee, he should have 0. So, perhaps, when he considers both members of her credal committee, he should be represented as having the credal set where he has both credence 1 and 0. So then

$$\mathrm{p}^{\mathrm{Owen}}(\neg \mathbb{P}_{\geqslant 1/2}^{\mathrm{Ann}}(Heads)) = \{0, 1\}$$

is the appropriate state.

An interesting feature of this is that Owen becomes an imprecise reasoner in virtue of reasoning about an imprecise agent.

In our probabilistic modal structures, as we have used them so far, all the worlds have a precise interpretation of the non-semantic states of affairs, as given by $\mathbf{M}$, and all the accessibility measures, m_w^{A}, are precise finitely additive probability measures. In that case imprecision can only come in with the semantic facts, so we wouldn't be able to have a situation where "Ann has degrees of belief with $\mathrm{p}^{\mathrm{Ann}}(Heads) = [0, 1]$, and Owen knows this" because Ann will always have a precise opinion about the chance of *Heads*. But we might be interested in such a situations and this is indeed the kind of case that is considered in the imprecise probability literature. We can build that possibility in by allowing imprecise probabilistic modal structures. These are very closely related to imprecise type-spaces, as considered in Ahn (2007).[4]

---

[4] One might also consider variants of probabilistic modal structures that allow the model of the non-semantic language to be supervaluational and assign a set of permissible interpretations. This would then allow one to model agent's attitudes towards vague states of affairs. We will not discuss that just to keep the chapter more focused.

Definition 4.2.2. An *imprecise probabilistic modal structure* is given by:

- W a non-empty set,

- $\mathbf{M}$ which assigns to each world w a model of $\mathcal{L}$, $\mathbf{M}(w)$,

- For each notion of probability A, an imprecise accessibility measure, $\mathbf{m}^{A}$, which is a collection of finitely additive probability measures over W.

Let $\mathfrak{M}$ be an imprecise probabilistic modal structure. If $m^{A} \in \mathbf{m}^{A}$ for each $A \in \mathsf{Agents}$, we define $\mathfrak{M}^{\langle m^{A} \rangle}$ be the precise probabilistic modal structure which has W and $\mathbf{M}$ as in $\mathfrak{M}$, but the accessibility measure given by m^{A} for each agent A.

Consider the interactive Ellsberg urn from Ahn (2007).

Example 4.2.3. Start with the Ellsberg setup (Ellsberg, 1961).

> I have an urn that contains ninety marbles. Thirty marbles are yellow. The remainder are blue or red in some unknown proportion.

With the additional following setup:

> Now suppose that we have two subjects, [Rosie] and [Billy], and the experimenter gives each an additional piece of information about the drawn ball. [Rosie] will be told if the ball is [red] or not, while [Billy] will be told if the ball is [blue] or not. (Ahn, 2007, names and colours altered).

This situation can be represented by the imprecise probabilistic modal structure represented in Fig. 4.2.
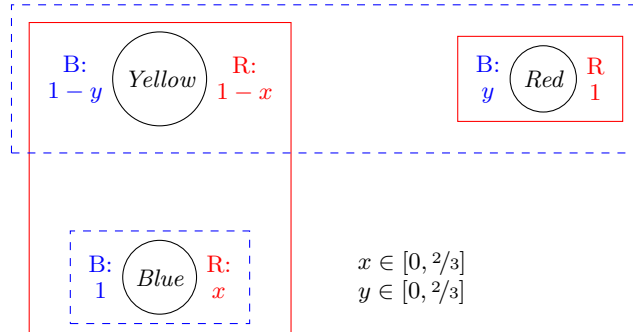
Figure 4.2: The red (solid) information on this diagram captures Rosie's accessibility measure, blue (dashed) information captures Billy's.

We can now present a probabilistic modal structure with the features of Example 4.2.1.

Example 4.2.4. We want to find a probabilistic modal structure with the features: Ann has degrees of belief with $\mathrm{p}^{\mathrm{Ann}}(Heads) = \{0,1\}$, and Owen knows this.

This doesn't give us information about what degree of belief Owen assigns to *Heads*, and how we fix that doesn't matter for what we are considering, so let's suppose that he knows whether coin landed heads or not.

We can then represent this by the probabilistic modal structure where $W = \{w_{Heads}, w_{Tails}\}$, $\mathbf{M}$ is as expected,[5] $\mathbf{m}^{\text{Ann}}$ is the collection of measure functions over $\{w_{Heads}, w_{Tails}\}$ with

$$m_w^{\text{Ann}}(\{w_{Heads}\}) = m_{w_{Tails}}^{\text{Ann}}(\{w_{Heads}\}) \in \{0, 1\},$$

and $\mathbf{m}^{\text{Owen}}$ is a singleton where its only member has the properties:

$$m_{w_{Heads}}^{\text{Owen}}(\{w_{Heads}\}) = m_{w_{Tails}}^{\text{Owen}}(\{w_{Tails}\}) = 1.$$

We can represent this imprecise probabilistic modal structure as in Fig. 4.3
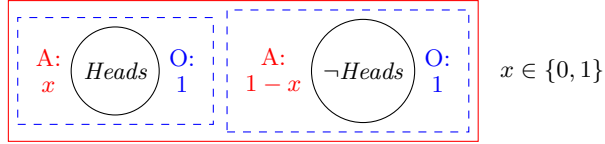
Figure 4.3: The red (solid) information on this diagram captures Ann's accessibility measure, the blue (dashed) information captures Owen's.

This has just set up the model for Example 4.2.1, in Example 4.2.7 we will analyse how to determine $p^{\text{Owen}}(\neg \mathbb{P}_{\geq 1/2}^{\text{Ann}}(Heads))$ formally. We do that by means of Θ analogous to Definition 4.1.2 but which now takes imprec-prob-eval functions to imprec-prob-eval functions.

The definition of Θ in Definition 4.1.2 was dependent on the probabilistic modal structure by virtue of being dependent on the definition of $(\mathbf{M}, \mathbf{p})(w) \models_{\mathfrak{M}} \varphi$ and we haven't yet defined

$$(\mathbf{M}, \mathbf{p})(w) \models_{\mathfrak{M}} \varphi$$

for $\mathfrak{M}$ an *imprecise* probabilistic modal structure. We defined $\Theta_{\mathfrak{M}}(\mathcal{P})$ to be the collection of $\Theta_{\mathfrak{M}}(\mathbf{p})$, using the precise version to directly get an imprecise version, and we can do the same sort of thing here. We define $\Theta_{\mathfrak{M}}$ by working with $\Theta_{\mathfrak{M}\langle m^{\text{Ag}}\rangle}$ on the precisifications of the probabilistic modal structure and then collecting the results.

Definition 4.2.5. Let $\mathfrak{M}$ be an imprecise probabilistic modal structure. Define

$$\Theta_{\mathfrak{M}}(\mathcal{P}) := \{\Theta_{\mathfrak{M}\langle m^{\text{A}}\rangle}(\mathbf{p}) \mid \mathbf{p} \in \mathcal{P} \text{ and for each A} \in \mathsf{Agents}, m^{\text{A}} \in \mathbf{m}^{\text{A}}\}$$

Where
$$\Theta_{\mathfrak{M}\langle m^{\text{A}}\rangle}(\mathbf{p})^{\text{A}}(w)(\varphi) = m_w\{v \mid (\mathbf{M}, \mathbf{p})(v) \models \varphi\}.$$

As always we will generally drop the explicit reference to $\mathfrak{M}$.

Proposition 4.2.6. Θ *is monotone, so it has fixed points. Fixed points will be called* stable states.

---

[5]I.e. $\mathbf{M}(w_{Heads}) \models Heads$, $\mathbf{M}(w_{Tails}) \models \neg Heads$.

Example 4.2.7 (Example 4.2.1 formalised)**.** Consider the probabilistic modal structure from Example 4.2.4. Let $\boldsymbol{\mathcal{P}}$ be any imprec-prob-eval function. Let m^{Owen} be the unique member of $\mathbf{m}^{\text{Owen}}$. Take any $\boldsymbol{\mathcal{P}}$.

We will first show that

$$\Theta_{\mathfrak{M}}(\boldsymbol{\mathcal{P}})(w)^{\text{Ann}}(\textit{Heads}) = \{0.1, 0.9\}.$$

And then

$$\Theta_{\mathfrak{M}}(\boldsymbol{\mathcal{P}})(w)^{\text{Owen}}(\neg\mathbb{P}^{\text{Ann}}_{\geqslant 1/2}(\textit{Heads})) = \{0, 1\}.$$

Consider $m^{0.1} \in \mathbf{m}^{\text{Ann}}$ with $m_w^{0.1}(\{w_{\textit{Heads}}\}) = 0.1$. Let $\Theta_{0.1}$ be a shorthand for $\Theta_{\mathfrak{M}\langle m^{0.1}, m^{\text{Owen}}\rangle}$. Let $\mathbf{p} \in \boldsymbol{\mathcal{P}}$.

$$\Theta_{0.1}(\mathbf{p})(w)^{\text{Ann}}(\textit{Heads}) = m_w^{0.1}\{v \mid (\mathbf{M}, \mathbf{p})(v) \models \textit{Heads}\}$$

And $(\mathbf{M}, \mathbf{p})(v) \models \textit{Heads} \iff \mathbf{M}(v) \models \textit{Heads}$, so:

$$= m_w^{0.1}\{w_{\textit{Heads}}\}$$
$$= 0.1$$

And therefore

$$(\mathbf{M}, \Theta_{0.1}(\mathbf{p}))(w) \models \neg\mathbb{P}^{\text{Ann}}_{\geqslant 1/2}(\textit{Heads})$$

for w both $w_{\textit{Heads}}$ and $w_{\textit{Tails}}$.

Consider $m^{0.9} \in \mathbf{m}^{\text{Ann}}$ with $m_w^{0.9}(\{w_{\textit{Heads}}\}) = 0.9$. Let $\Theta_{0.9}$ be a shorthand for $\Theta_{\mathfrak{M}\langle m^{0.9}, m^{\text{Owen}}\rangle}$ we can apply directly analogous reasoning. We get that

$$(\mathbf{M}, \Theta_{0.9}(\mathbf{p}))(w) \models \mathbb{P}^{\text{Ann}}_{\geqslant 1/2}(\textit{Heads})$$

for w both $w_{\textit{Heads}}$ and $w_{\textit{Tails}}$.

Now,

$$\Theta_{\mathfrak{M}}(\boldsymbol{\mathcal{P}}) = \{\Theta_{0.1}(\mathbf{p}) \mid \mathbf{p} \in \boldsymbol{\mathcal{P}}\} \cup \{\Theta_{0.9}(\mathbf{p}) \mid \mathbf{p} \in \boldsymbol{\mathcal{P}}\}$$

So,

$$\Theta_{\mathfrak{M}}(\boldsymbol{\mathcal{P}})(w)^{\text{Ann}}(\textit{Heads}) = \{0.1, 0.9\}.$$

Now

$$\Theta_{\mathfrak{M}}(\Theta_{\mathfrak{M}}(\boldsymbol{\mathcal{P}})) = \{\Theta_{\mathfrak{M}\langle m^{\text{A}}\rangle}(\mathbf{p}') \mid \mathbf{p}' \in \Theta_{\mathfrak{M}}(\boldsymbol{\mathcal{P}}) \text{ and for each A} \in \mathsf{Agents},\, m^{\text{A}} \in \mathbf{m}^{\text{A}}\}$$
$$= \{\Theta_{\mathfrak{M}\langle m^{\text{A}}\rangle}(\Theta_{0.1}(\mathbf{p})) \mid \mathbf{p} \in \boldsymbol{\mathcal{P}} \text{ and for each A} \in \mathsf{Agents},\, m^{\text{A}} \in \mathbf{m}^{\text{A}}\}$$
$$\cup \{\Theta_{\mathfrak{M}\langle m^{\text{A}}\rangle}(\Theta_{0.9}(\mathbf{p})) \mid \mathbf{p} \in \boldsymbol{\mathcal{P}} \text{ and for each A} \in \mathsf{Agents},\, m^{\text{A}} \in \mathbf{m}^{\text{A}}\}$$
$$= \{\Theta_{0.1}(\Theta_{0.1}(\mathbf{p})) \mid \mathbf{p} \in \boldsymbol{\mathcal{P}}\} \cup \{\Theta_{0.9}(\Theta_{0.1}(\mathbf{p})) \mid \mathbf{p} \in \boldsymbol{\mathcal{P}}\}$$
$$\cup \{\Theta_{0.1}(\Theta_{0.9}(\mathbf{p})) \mid \mathbf{p} \in \boldsymbol{\mathcal{P}}\} \cup \{\Theta_{0.9}(\Theta_{0.9}(\mathbf{p})) \mid \mathbf{p} \in \boldsymbol{\mathcal{P}}\}$$

For x either 0.1 or 0.9, we have:

$$\Theta_x(\Theta_{0.1}(\mathbf{p}))(w)^{\text{Owen}}(\neg\mathbb{P}^{\text{Ann}}_{\geqslant 1/2}\textit{Heads})$$
$$= m_w^{\text{Owen}}\{v \mid (\mathbf{M}, \Theta_{0.1}(\mathbf{p}))(w) \models \neg\mathbb{P}^{\text{Ann}}_{\geqslant 1/2}(\textit{Heads})\}$$
$$= m_w^{\text{Owen}}(W) = 1$$

and

$$\Theta_x(\Theta_{0.9}(\mathbf{p}))(w)^{\text{Owen}}(\neg\mathbb{P}^{\text{Ann}}_{\geqslant 1/2}Heads)$$
$$= m_w^{\text{Owen}}\{v \mid (\mathbf{M}, \Theta_{0.9}(\mathbf{p}))(w) \models \neg\mathbb{P}^{\text{Ann}}_{\geqslant 1/2}(Heads)\}$$
$$= m_w^{\text{Owen}}(\varnothing) = 0$$

Therefore

$$\Theta_{\mathfrak{M}}(\Theta_{\mathfrak{M}}(\boldsymbol{\mathcal{P}}))(w)^{\text{Owen}}(\neg\mathbb{P}^{\text{Ann}}_{\geqslant 1/2}(Heads)) = \{0, 1\}.$$

We can also observe that this is stable with respect to Owen's beliefs about $\neg\mathbb{P}^{\text{Ann}}_{\geqslant 1/2}(Heads)$. So stable evaluations $\boldsymbol{\mathcal{P}}$ will have

$$\boldsymbol{\mathcal{P}}(w_{Heads})^{\text{Owen}}(\neg\mathbb{P}^{\text{Ann}}_{\geqslant 1/2}) = \{0, 1\}.$$

This is as our informal description worked. The additional complication over the intuitive explanation is just to allow for generality.

4.3 Convexity?

It is sometimes argued that imprecise credal states should be convex in order to be behaviouristically distinct because we can capture lower and upper probabilities via buying and selling behaviour. This argument can be avoided if one takes conditional bets, but we do not do this because of the issues suggested in Section 1.7. However, there are other reasons we may not want to consider only convex sets: for example, we might want to consider an agent to believe two paramaters to be probabilistically independent, and this cannot be represented by a convex credal committee.

If one did want to regain the convexity, one can always take the convex closure of what has been done here. Alternatively one could generalise the work on the strong Kleene Kripkean semantics more simply and just work with $\overline{\mathrm{p}_{(w,f)}}$ and $\underline{\mathrm{p}_{(w,f)}}$ as in Definition 3.2.15. This would work as one might expect.

However, we will loose something by taking convex sets: it will no longer be the case that every member of the credal state looks best from some (possibly different) member's perspective. For example in the case of π in $\mathfrak{M}_{\text{omn}}$ (Example 4.1.7) the only probability values for π that can look best from any prob-eval function's perspective are 0 and 1, and moreover they both have to be in a stable state, so the stable state in that situation is not convex. If instead we considered some convex set we would have that every member of the credal state is in the convex closure of the credal states which look best from some member's perspective. But that is much less appealing. We therefore instead suggest that one works with these non-convex sets.

Appendix 4.A   Using evaluation functions

Consider evaluation functions as in Definition 3.2.2 which work as extensions of the truth predicate. If p is maximally consistent, then it determines a classical model of $\mathcal{L}_{\mathsf{P},\mathsf{T}}$ at each world, and each P will be probabilistic. So, given a set of evaluation functions we have a set of interpretations of P and T that are

nicely coordinated. When considering supervaluational versions of the Kripke construction in the truth case, one usually works with a single evaluation function and determine a set from that, for example all the maximally consistent evaluation functions extending the single one, and from that determine using supervaluational logic which sentences will be super-true and use that as the next single evaluation function. However that way of doing things imposes the restriction that one can only consider sets of evaluation functions which are generated by a single evaluation function. We do not wish to impose this restriction because that would impose a restriction on the sets of probabilities that are admitted as the imprecise probabilities, for example it would require that we have intervals assigned to each sentence, and perhaps lead us to results about which sets are good that is more restrictive than is desired. We only focus on maximally consistent evaluation functions because this leads us to probability functions, so we are then working with sets of probability functions, which allows this work to fit well with work on imprecise probabilities.

To provide a Kripke construction as in Chapter 3, we need to provide some inductive definition giving a fixed point.

Definition 4.A.1. f is a maximally consistent evaluation function if for each w, $f(w)$ is a maximally consistent set of sentences (or the codes thereof).

An SV-evaluation, F, is a collection of maximally consistent evaluation functions.

If the evaluation function f is maximally consistent, then it determines a model of $\mathcal{L}_{\mathsf{P},\mathsf{T}}$ at each world where each P is probabilistic. This can be defined as a classical model using $\mathbf{M}(w)$ to interpret the vocabulary from $\mathcal{L}$, and defining $(w, f) \models_{\mathfrak{M}} \mathsf{P}^{\ulcorner}\varphi^{\urcorner} = r \iff m_w\{v \mid \#\varphi \in f(v)\} = r$, which corresponds to that in Definition 3.2.3 for when f is maximally consistent.

Definition 4.A.2. Define $\Theta(f)$ by:

$$\#\varphi \in \Theta(f)(w) \iff (w, f) \models_{\mathfrak{M}} \varphi$$

For F an SV-evaluation, define

$$\Theta(F) := \{\Theta(f) \mid f \in F\}.$$

This is monotone so fixed points exist. We should have

Conjecture 4.A.3. $\mathcal{P}$ *is a fixed point iff there is a fixed point SV-evaluation* F *such that* $\mathbf{p} \in \mathcal{P}$ *iff there is* $f \in F$ *such that*

$$\mathbf{p}(w)(\varphi) = m_w\{v \mid \varphi \in f(v)\}.$$

Which would show these constructions are essentially the same.

f

Chapter 5

The Revision Theory of Probability

In the previous chapters we have developed semantics which drop certain traditional probability axioms and assumptions. In this chapter we will consider an alternative theory of probability where we have that the standard probability axioms are retained.

One influential theory for the liar paradox is the revision theory of truth. The revision theory of truth was independently developed by Gupta and Herzberger and the idea is to improve, stage-by-stage, some arbitrary model of the language. Unlike for the Kripkean construction, such a construction will never terminate but will instead result in a transfinite sequence of models. This lack of a "fixed point" is the price to pay for remaining fully classical. In this chapter we see how one can develop a revision construction for probability. Since the underlying logic is fully classical our probability notion will satisfy the usual axioms of probability (at least for finitely additive probability). We can therefore work with a language with a probability function symbol (see Definition 1.6.11).

We shall present two different revision constructions for this language. In the first construction we will develop Leitgeb's work from Leitgeb (2008, 2012). This construction cannot apply to general interpretations of probability but instead fixes it to something that might be considered as semantic probability. The second will be based on possible world structures and can be used to give a theory for probabilities in the form of subjective probabilities or objective chances. This is because it is based on background probabilistic modal structures.

One of our aims in developing our constructions is to provide nice limit stages that can themselves be used as good models for probability and truth. To achieve this we shall provide a strong criterion for what the limit models should look like by giving a strong way of precisifying the claim:

> If a property of interest of the interpretations is brought about by the sequence beneath μ then it should be satisfied at the μth stage.

In Gupta and Belnap's *locus classicus* on revision theories (Gupta and Belnap, 1993) the authors just consider the properties "φ is true" and "φ is false", and understand "brought about" according to what they call a stability condition. Note that this notion of stability is not connected to that in Chapter 4. For

Gupta and Belnap, if φ is true stably beneath μ, meaning that from some point onwards, φ is always true, then φ should also be true at the stage μ; and similarly for falsity. This is a weak way to make this criterion precise and they show that even such a weak characterisations leads to an interesting construction. We will instead present a strong limit stage criterion, the particular change being that we consider more properties. For example we will also be interested in properties like

> The probability of φ is equal to the probability of ψ.

It is interesting to study strong proposals as well as weak proposals because from strong proposals we can obtain nice models at the limit stages which may have different kinds of properties to the models obtainable at the successor stages.

Outline of the chapter

In more detail what to come is as follows. Firstly in Section 5.1 we present some definitions we will use throughout the chapter. The chapter properly begins with Section 5.2 where we shall present our revision construction that extends Leitgeb's. This relies on the notion of relative frequencies to interpret the probability notion at the successor stages. At limit stages we shall present a criterion which is stronger than the one usually considered. We will be interested in any properties of interpretations that act nicely at limit stages. For example, a property described by

$$\mathrm{p}(\varphi) \leqslant 1/2$$

will act nicely at limits, whereas

$$\mathrm{p}(\varphi) < 1/2$$

will not. If some such property is "brought about" by the previous stages, in the sense of being *nearly stable* beneath μ (for μ a limit ordinal), then we ask that it is satisfied at the stage μ. We shall motivate and present the construction and in Example 5.2.9 we discuss some examples of how it works. In Section 5.2.2 we shall show some properties of the construction. This includes the property that at every state P is interpreted as a finitely additive probability function and that at limit stages P and T satisfy the so-called Probabilistic Convention T. We then present some alternatives in Section 5.2.3 which weaken the limit constraint, before moving on to comments on this style of construction in Section 5.2.4. A major feature of the construction is that it seems hard to use these constructed interpretations for standard uses of probability such as to model an agent's beliefs.

In Section 5.3 we develop an alternative construction that can be used to model subjective probabilities. This works by using an underlying structure in the form of a probabilistic modal structure to give extra information about how the probability notion should work. We discuss how the limit stage should be defined in Sections 5.3.2 and 5.3.3. Then finally in Section 5.4 we present some axiomatic theories for all these constructions. These have nothing like the completeness that we obtained in Chapter 3. We obtain a nice theory and nice properties if we just focus on limit stages. This is because of our fairly restrictive definition for the limit stages, and it is contrary to traditional revision

constructions. For example at a limit state we will have

$$\neg\varphi \in \mathrm{T} \iff \varphi \notin \mathrm{T}.$$

We close the chapter with a short conclusions section.

5.1 Preliminaries

Setup 5 (for Chapter 5). *Let $\mathcal{L}$ be some language extending $L_{\mathsf{PA,ROCF}}$ containing the predicates N and R. In this chapter we will be working with $\mathcal{L}_{\mathsf{P,T}}$ as presented in Definition 1.6.11 which extends $\mathcal{L}$ with a function symbol P and a unary predicate T.*

We can use this more flexible language instead of $\mathcal{L}_{\mathsf{P}_{\geqslant},\mathsf{T}}$ that was used in Chapter 3 because our semantics will be based on classical logic and we will assign single point valued probabilities to each sentence.

Note that in the specification of this language we required that $\mathcal{L}$ be countable in order to still use arithmetic for our coding. This assumption isn't essential to anything that is done in this chapter.

Notation 5.1.1. Models of the language $\mathcal{L}_{\mathsf{P,T}}$ shall be denoted by $\mathcal{M} = (\mathrm{M}, \mathrm{T}, \mathrm{p})$. Here, M is some background model of $\mathcal{L}$, which we assume to interpret the $L_{\mathsf{PA,ROCF}}$ vocabulary as intended, for example interpreting N and R by the set of (standard) natural and real numbers, respectively. T is a subset of $\mathsf{Sent}_{\mathsf{P,T}}$ which interprets the predicate T. p is a function from $\mathsf{Sent}_{\mathsf{P,T}}$ to $\mathbb{R}$ which provides the interpretation of the function symbol P.[1]

Revision sequences will be (transfinite) sequences of models of $\mathcal{L}_{\mathsf{P,T}}$, the αth model of which is denoted $\mathcal{M}_\alpha = (\mathrm{M}, \mathrm{T}_\alpha, \mathrm{p}_\alpha)$.

Let $[\![\varphi]\!]_{\mathcal{M}} := \begin{cases} 1 & \mathcal{M} \models \varphi \\ 0 & \text{otherwise} \end{cases}$.

In this chapter we will often identify a sentence with its code, but this should not lead to any confusion.

5.2 Revising probability using relative frequencies and near stability

5.2.1 Motivating and defining the revision sequence

The final definition of the construction can be found in Definition 5.2.8 on Page 127 but instead of jumping straight into giving the definitions we will explain how it is motivated.

The idea of a revision construction is to start with some (classical) model of $\mathcal{L}_{\mathsf{P,T}}$ and to improve on it. To do this we need to pick a sequence of extensions of T and P, i.e. develop a sequence $\mathcal{M}_\alpha$ of interpretations of T and P.

---

[1] Here we are identifying a sentence with its code. More carefully we have that T provides the interpration for T by having $n \in \mathsf{T}^{\mathcal{M}} \iff$ there is some $\varphi \in \mathrm{T}$ with $n = \#\varphi$. Similarly for p.

Consider the liar sentence λ, which is a sentence where

$$\lambda \leftrightarrow \neg \mathsf{T}\ulcorner\lambda\urcorner$$

is arithmetically derivable. At the zero$^{\text{th}}$ stage suppose we have taken $\lambda \notin \mathrm{T}_0$. Then this zero$^{\text{th}}$ model will in fact satisfy $\neg \mathsf{T}\ulcorner\lambda\urcorner$, i.e. it will satisfy λ. At the next stage we therefore say λ is true, so put $\lambda \in \mathrm{T}_1$. By continuing this reasoning we see that the liar sentence will continue to flip in and out of the extension of the truth predicate:

$$
\begin{array}{cccccc}
\overset{\bullet}{\neg \mathsf{T}\ulcorner\lambda\urcorner} & \overset{\bullet}{\mathsf{T}\ulcorner\lambda\urcorner} & \overset{\bullet}{\neg \mathsf{T}\ulcorner\lambda\urcorner} & \overset{\bullet}{\mathsf{T}\ulcorner\lambda\urcorner} & \overset{\bullet}{\neg \mathsf{T}\ulcorner\lambda\urcorner} & \dashrightarrow \\
\lambda & \neg\lambda & \lambda & \neg\lambda & \lambda &
\end{array}
$$

A formal characterisation of this reasoning is given by the clause:

$$\varphi \in \mathrm{T}_{n+1} \iff \mathcal{M}_n \models \varphi$$

To develop a revision sequence for probability we also need to interpret p_n. Leitgeb (2012) presents a model of $\mathcal{L}_{\mathsf{P},\mathsf{T}}$ which he uses to prove the consistency of a set of principles including Probabilistic Convention T. To construct his model he first develops an ω-length sequence of models of $\mathcal{L}_{\mathsf{P},\mathsf{T}}$, which we can take to be the finite stages of a revision construction, the final model he proposes is something like a limit of the other models and can be seen as the ω-stage of a revision construction. We shall use his proposal to tell us how to interpret the finite stage probabilities and although our limit stage will differ in technical details it results in something very close to Leitgeb's ω-stage definition. Leitgeb's finite stage construction says that we should interpret the nth probability of φ as the relative frequency of φ being satisfied in the sequence of models leading up to n, i.e.,

$$\mathrm{p}_{n+1}(\varphi) = \frac{[\![\varphi]\!]_{\mathcal{M}_0} + \ldots + [\![\varphi]\!]_{\mathcal{M}_n}}{n+1}$$

Including probability into the diagram of the revision sequence for the liar sentence we then get:

$$
\begin{array}{cccccc}
\overset{\bullet}{\neg \mathsf{T}\ulcorner\lambda\urcorner} & \overset{\bullet}{\mathsf{T}\ulcorner\lambda\urcorner} & \overset{\bullet}{\neg \mathsf{T}\ulcorner\lambda\urcorner} & \overset{\bullet}{\mathsf{T}\ulcorner\lambda\urcorner} & \overset{\bullet}{\neg \mathsf{T}\ulcorner\lambda\urcorner} & \dashrightarrow \\
\lambda & \neg\lambda & \lambda & \neg\lambda & \lambda & \\
 & \mathsf{P}\ulcorner\lambda\urcorner = 1 & \mathsf{P}\ulcorner\lambda\urcorner = \nicefrac{1}{2} & \mathsf{P}\ulcorner\lambda\urcorner = \nicefrac{2}{3} & \mathsf{P}\ulcorner\lambda\urcorner = \nicefrac{2}{4} & \\
 & & & & = \nicefrac{1}{2} &
\end{array}
$$

However our job is not yet done. Although each new model is an improvement on the previous model, each of them is still in need of improvement. We therefore should extend the revision sequence to the transfinite. This is also required, for example, to assign appropriate truth values to some unproblematic quantified sentences like, e.g.

$$\forall n \in N(\mathsf{T}^n \ulcorner 0 = 0 \urcorner).$$

Once the limit stages have been defined our job will still not be done because that limit stage will itself be in need of improvement. We therefore will also have to define transfinite successor stages and work our way all the way up the ordinals. We first consider how the transfinite successor stages should be defined as before considering the limit stage.

5.2 Relative frequencies and near stability

The transfinite successor stages for truth are usually defined in the same way as the finite successor stages, namely

$$\varphi \in \mathrm{T}_{\alpha+1} \iff \mathcal{M}_\alpha \models \varphi$$

Following this same idea we would want to define $\mathrm{p}_{\alpha+1}(\varphi)$ to be the relative frequency of φ being satisfied in the sequence of models up to $\alpha + 1$. However there are now transfinitely-many such models so how should one define relative frequency in this case? We shall define $\mathrm{p}_{\alpha+1}(\varphi)$ to be the relative frequency of φ being satisfied in the finite sequence of models just before $\alpha + 1$.

For example,

$$\mathrm{p}_{\omega+5}(\varphi) = \frac{[\![\varphi]\!]_{\mathcal{M}_\omega} + [\![\varphi]\!]_{\mathcal{M}_{\omega+1}} + [\![\varphi]\!]_{\mathcal{M}_{\omega+2}} + [\![\varphi]\!]_{\mathcal{M}_{\omega+3}} + [\![\varphi]\!]_{\mathcal{M}_{\omega+4}}}{5}$$

For the liar sentence one would then obtain the following probabilities.

To give this definition in general one needs to choose the finite sequence of models beneath an ordinal.

Definition 5.2.1. For a successor ordinal $\alpha + 1$ we let $\zeta_{\alpha+1}$ denote the greatest limit ordinal beneath $\alpha + 1$, and $k_{\alpha+1}$ be the natural number such that $\alpha + 1 = \zeta_{\alpha+1} + k_{\alpha+1}$.

For example $\zeta_{\omega+5} = \omega$ and $k_{\omega+5} = 5$.

Proposition 5.2.2. $\zeta_{\alpha+n} = \zeta_\alpha$ for $n \in \omega$. And $k_{\alpha+n} = k_\alpha + n$.

We can define:

$$\mathrm{p}_{\alpha+1}(\varphi) = \frac{[\![\varphi]\!]_{\mathcal{M}_{\zeta_\alpha}} + [\![\varphi]\!]_{\mathcal{M}_{\zeta_\alpha+1}} + \ldots + [\![\varphi]\!]_{\mathcal{M}_{\zeta_\alpha+k_\alpha-1}} + [\![\varphi]\!]_{\mathcal{M}_\alpha}}{k_{\alpha+1}}$$

Now we can move on to approach the question of how to define the limit stages. In Gupta's first published presentation of his theory he says:

> Intuitively what is wanted is a way of summing up the improvements that are brought about by each successive application of τ_M [the function from $\mathcal{M}_\alpha$ to $\mathcal{M}_{\alpha+1}$[2]] That is, we want a way of going from the improvements that are severally brought about by the various applications of τ_M to the improvements that are collectively brought about by those applications. (Gupta, 1982, p. 39)

We might characterise this intended limiting procedure as:

> If a property of interest of the interpretations of T is "brought about" by the sequence beneath μ then it should be satisfied at the μth stage.

---

[2]Observe that in our construction no such single function is available since the $\alpha + 1^{\text{th}}$ stage depends on a number of stages beneath $\alpha + 1$.

This has to be filled-out by explaining which properties are of interest and what counts as being "brought about by the sequence beneath μ".

Gupta proposes that the properties which we should consider are those of the form

φ is in the extension of T

and

φ is not in the extension of T.

He characterises what is "brought about by the sequence beneath μ" by using the notion of *stability*: a stable property is one which at some point becomes satisfied and remains satisfied in the later improvements (for a detailed definition of stability see Definition 5.2.7).

This limit definition is then formalised by:

If φ is in the extension of T stably beneath μ, then $\varphi \in \mathsf{T}_\mu$, and similarly for φ not being in the extension of T.

Here is an example of how this works: suppose $\mathsf{T}_0 = \varnothing$, then:

0	1	2		ω	$\omega + 1$	
$\bullet$	$\bullet$	$\bullet$	$\dashrightarrow$	$\bullet$	$\bullet$	$\dashrightarrow$
$0 = 0$	$0 = 0$	$0 = 0$		$0 = 0$	$0 = 0$	
$\neg\mathsf{T}\ulcorner 0 = 0\urcorner$	$\mathsf{T}\ulcorner 0 = 0\urcorner$	$\mathsf{T}\ulcorner 0 = 0\urcorner$		$\mathsf{T}\ulcorner 0 = 0\urcorner$	$\mathsf{T}\ulcorner 0 = 0\urcorner$	

From stage 1 onwards, $0 = 0$ is in the extension of T, i.e. $0 = 0 \in \mathsf{T}_n$ for $n \geqslant 1$, so the stability requirement will therefore lead us to that $0 = 0 \in \mathsf{T}_\omega$.

We shall propose a stronger way to precisify the intended limiting procedure. Firstly we can consider more properties of the interpretations. Secondly we shall extend the definition of what is to be counted as a property being brought about.

We shall allow more properties to be considered than just the ones that take the form $\varphi \in \mathsf{T}$ and $\varphi \notin \mathsf{T}$. We would ideally allow any properties to be considered but are unable to do this.[3] To see this consider the probability of $\mathsf{T}\ulcorner 0 = 0\urcorner$ starting with an initial model where $p_0(\mathsf{T}\ulcorner 0 = 0\urcorner) = 0$ and $\mathsf{T}_0 = \varnothing$:

0	1	2	3	
$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\dashrightarrow$
$0 = 0$	$0 = 0$	$0 = 0$	$0 = 0$	
$\neg\mathsf{T}\ulcorner 0 = 0\urcorner$	$\mathsf{T}\ulcorner 0 = 0\urcorner$	$\mathsf{T}\ulcorner 0 = 0\urcorner$	$\mathsf{T}\ulcorner 0 = 0\urcorner$	
$p_0(\mathsf{T}\ulcorner 0 = 0\urcorner) = 0$	$p_1(\mathsf{T}\ulcorner 0 = 0\urcorner) = 0$	$p_2(\mathsf{T}\ulcorner 0 = 0\urcorner) = \nicefrac{1}{2}$	$p_3(\mathsf{T}\ulcorner 0 = 0\urcorner) = \nicefrac{2}{3}$	

We can also graph these probability values as follows:

---
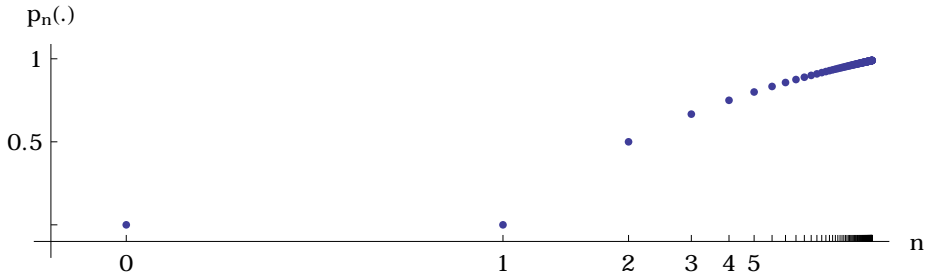
[3]Seamus Bradly motivated me to consider which properties can be worked with leading me to this characterisation of the limit stages.

5.2 Relative frequencies and near stability

For each $\epsilon \in \mathbb{R}^{>0}$, $1 > \mathrm{p}(\mathsf{T}\ulcorner 0 = 0 \urcorner) > 1 - \epsilon$ is stable in the sequence beneath ω, so it counts as being brought about by that sequence. If we consider each of these properties as being applicable in the limiting procedure we would ask each of them to be satisfied, i.e. for each $\epsilon > 0$, $1 > \mathrm{p}_\omega(\mathsf{T}\ulcorner 0 = 0 \urcorner) > 1 - \epsilon$. However this cannot be satisfied because there is no value of $\mathrm{p}_\omega(\mathsf{T}\ulcorner 0 = 0 \urcorner)$ which will satisfy this constraint *for all* ϵ. The difference is that $\{x \mid r < x < q\}$ is an open set, but $\{x \mid r \leqslant x \leqslant q\}$ is closed. Closed sets are the ones which contain their limit points, so one can only require nice limiting behaviour with respect to closed sets. We transfer the notion of such closed sets directly to requirements on the models. To do this we first need to put a topology on the space of models.

Definition 5.2.3. Mod denotes the set of all models of $\mathcal{L}_{\mathsf{T},\mathsf{P}}$. Remember these all take the form $(\mathrm{M}, \mathrm{p}, \mathrm{T})$.

Take as a subbase for the closed sets all sets of the form:

- for each $\mathrm{M}_0 \in \mathsf{Mod}_\mathcal{L}$, $\{(\mathrm{M}, \mathrm{p}, \mathrm{T}) \mid \mathrm{M} \neq \mathrm{M}_0\}$,

- $\varphi \in \mathrm{T}$ and $\varphi \notin \mathrm{T}$,
  i.e. $\{(\mathrm{M}, \mathrm{p}, \mathrm{T}) \mid \varphi \in \mathrm{T}\}$ and $\{(\mathrm{M}, \mathrm{p}, \mathrm{T}) \mid \varphi \notin \mathrm{T}\}$,[4]

- $\mathrm{p}(\varphi) \leqslant r$ and $\mathrm{p}(\varphi) \geqslant r$ for $r \in \mathbb{R}$,
  i.e. $\{(\mathrm{M}, \mathrm{p}, \mathrm{T}) \mid \mathrm{p}(\varphi) \leqslant r\}$ and $\{(\mathrm{M}, \mathrm{p}, \mathrm{T}) \mid \mathrm{p}(\varphi) \geqslant r\}$.

So using this subbase we can give an inductive definition of being closed: $C \subseteq \mathsf{Mod}$ *is closed* if

- C is a member of the subbase for closed sets,

- $C = A \cup B$ for some A, B closed,

- $C = \bigcap_{i \in I} A_i$ with each A_i closed.

For $\varphi \in \mathsf{Sent}_{\mathsf{P},\mathsf{T}}$ we say φ *is closed* if $[\varphi] = \{\mathcal{M} \in \mathsf{Mod} \mid \mathcal{M} \models \varphi\}$ is closed.[5]

The topology is defined so that Mod is topologically equivalent to $\mathsf{Mod}_\mathcal{L} \times \mathbb{R}^{\mathsf{Sent}_{\mathsf{P},\mathsf{T}}} \times \{0,1\}^{\mathsf{Sent}_{\mathsf{P},\mathsf{T}}}$ with the product topology (where $\mathsf{Mod}_\mathcal{L}$ is endowed with the discrete topology). Thus, any set which does not depend on T or p is closed. Ignoring the part involving $\mathcal{L}$, this is the sort of topology used in Christiano et al. (ms).

The following states a useful way of checking whether a set of models is closed or not.

Proposition 5.2.5. *The following are equivalent:*

---

[4]We will often identify the property, i.e. $\varphi \in \mathrm{T}$, with the set $\{(\mathrm{M}, \mathrm{p}, \mathrm{T}) \mid \varphi \in \mathrm{T}\}$. We therefore apply our notion of closedness to subsets, characterisations of features of the models and sentences; but this should not cause any confusion.

[5]We conjecture:

Conjecture 5.2.4. $[\varphi]$ *is closed iff φ is equivalent to a sentence of the form*

$$\forall x_1 \dots x_n \bigwedge_{i \in I} \bigvee_{i \in J} \psi_{i,j}$$

where I and J are finite and $\psi_{i,j}$ takes one of forms of the members of the subbase.

1. $C \subseteq \mathsf{Mod}$ *is closed*

2. *For every sequence* $(M, \mathrm{p}_n, \mathrm{T}_n)$ *and each* $(M, \mathrm{p}_{lim}, \mathrm{T}_{lim})$ *such that* $(M, \mathrm{p}_n, \mathrm{T}_n) \longrightarrow (M, \mathrm{p}_{lim}, \mathrm{T}_{lim})$, *i.e.*

   - *for every* φ $\mathrm{p}_n(\varphi) \longrightarrow \mathrm{p}_{lim}(\varphi)$[6]
   - *for every* φ *there is some* $m \in \mathbb{N}$ *such that either*
     - *for every* $n > m$, $\varphi \in \mathrm{T}_n$ *and* $\varphi \in \mathrm{T}_{lim}$, *or*
     - *for every* $n > m$, $\varphi \notin \mathrm{T}_n$ *and* $\varphi \notin \mathrm{T}_{lim}$.

   If each $(M, \mathrm{p}_n, \mathrm{T}_n) \in C$, *then* $(M, \mathrm{p}_{lim}, \mathrm{T}_{lim}) \in C$.

In its current form the theorem is only true when the language is countable. If it is not countable, then the theorem would still hold with the proviso that $C \subseteq \mathsf{Mod}$ depend on the interpretation of probability and truth for only countably many sentences. The details of this restriction and the proof of this result can be found in Section 5.A.

Such closed properties will therefore be the ones that we can ask to be carried over to the limits. So the strongest possible interpretation of "the properties to be considered" is all the closed sets. This will lead to the limiting behaviour in the features described in Example 5.2.9. First we give some examples of sets which are closed and not.

Example 5.2.6. The following are closed:

- Any φ where $\varphi \in \mathsf{Sent}_{\mathcal{L}}$,

- $\mathrm{p}(\varphi) = \nicefrac{1}{2}$,

- $\mathrm{p}(\varphi) = 1 \vee \mathrm{p}(\varphi) = 0$,

- $\mathrm{p}(\varphi) - \mathrm{p}(\mathsf{T}\ulcorner\varphi\urcorner) \leqslant \epsilon$

- $\mathrm{p}(\varphi) = 1 - \mathrm{p}(\psi)$,

- $\mathrm{p}(\varphi) \cdot \mathrm{p}(\psi) \geqslant \nicefrac{1}{2}$,

- $\mathrm{p}(\psi) = 0$,

- If $\mathrm{p}(\varphi) > 0$, then $\mathrm{p}(\psi) = 0$ (because "if A then B" is "not-A or B")

- If $\mathrm{p}(\psi) > 0$, then $\dfrac{\mathrm{p}(\varphi \wedge \psi)}{\mathrm{p}(\psi)} = \nicefrac{1}{2}$,

- $\mathrm{p}(\psi) + \mathrm{p}(\chi) = \mathrm{p}(\psi \vee \chi)$,

- p is a finitely additive probability function,[7]

- Any φ which is in the propositional language containing the propositional variables $\mathsf{T}\ulcorner\psi\urcorner$ for all $\varphi \in \mathsf{Sent}_{\mathsf{P},\mathsf{T}}$, E.g.

  - $\mathsf{T}\ulcorner\neg\varphi\urcorner \leftrightarrow \neg\mathsf{T}\ulcorner\varphi\urcorner$,

---

[6]I.e. for all $\epsilon \in \mathbb{R}_{>0}$, there is some $m \in \mathbb{N}$ such that for all $n > m$, $|\mathrm{p}_n(\varphi) - \mathrm{p}_{\lim}(\varphi)| < \epsilon$.

[7]I became aware of the fact that the property of being a finitely additive probability function is closed in Christiano et al. (ms).

        $- \ (\mathsf{T}^\ulcorner \varphi_1{}^\urcorner \wedge \ldots \wedge \mathsf{T}^\ulcorner \varphi_n{}^\urcorner) \to \mathsf{T}^\ulcorner \psi^\urcorner,$

- T is maximally consistent.

The following are *not* closed:

- $\mathrm{p}(\varphi) \in \mathbb{Q}$,

- $^1\!/_2 < \mathrm{p}(\varphi)$,

- $\mathrm{p}(\varphi) \neq {}^1\!/_2$,

- $\neg\forall n{\in}N\,\mathsf{T}^{n+1}{}^\ulcorner \gamma^\urcorner$, or equivalently $\exists n{\in}N\,\neg\mathsf{T}^{n+1}{}^\ulcorner\gamma^\urcorner$,

- p is an $\mathbb{N}$-additive probability function,

- T is ω-consistent.

Proof. Most of these can either be seen directly or shown by using Proposition 5.2.5.

For example, consider $\mathrm{p}(\varphi) = 1 - \mathrm{p}(\psi)$. This can be shown either by the sequential characterisation or by reformulating it. Reformulated it would be:

$$\{\mathcal{M} \mid \mathrm{p}(\varphi) = 1 - \mathrm{p}(\psi)\}$$
$$= \bigcap_{\substack{r+q<1 \\ r,q\geqslant 0}} (\{\mathcal{M} \mid \mathrm{p}(\varphi) \geqslant r\} \cup \{\mathcal{M} \mid \mathrm{p}(\psi) \geqslant q\})$$
$$\cap \bigcap_{\substack{r+q>1 \\ r,q\geqslant 0}} (\{\mathcal{M} \mid \mathrm{p}(\varphi) \leqslant r\} \cup \{\mathcal{M} \mid \mathrm{p}(\psi) \leqslant q\})$$

To show that being finitely additive is closed we use the criterion of being sequentially closed. Take any sequence $(\mathrm{M}^n, \mathrm{p}^n, \mathrm{T}^n)$ with $\mathrm{p}^n(\psi) + \mathrm{p}^n(\chi) = \mathrm{p}^n(\psi \vee \chi)$. Suppose that for every φ, $\mathrm{p}^n(\varphi) \longrightarrow \mathrm{p}^{\lim}(\varphi)$. Then:

$$\mathrm{p}^n(\psi \vee \chi) \longrightarrow \mathrm{p}^{\lim}(\psi \vee \chi)$$
$$\mathrm{p}^n(\psi \vee \chi) = \mathrm{p}^n(\psi) + \mathrm{p}^n(\chi) \longrightarrow \mathrm{p}^{\lim}(\psi) + \mathrm{p}^{\lim}(\chi)$$

So in fact $\mathrm{p}^{\lim}(\psi \vee \chi) = \mathrm{p}^{\lim}(\psi) + \mathrm{p}^{\lim}(\chi)$. Therefore $(\mathrm{M}, \mathrm{p}^{\lim}, \mathrm{T}^{\lim}) \in C$.

One can easily observe that $\mathrm{p}(\varphi) \geqslant 0$ and $\mathrm{p}(\varphi) = 1$ are closed for any φ. So

$$C_{\text{fin add}} = \bigcap_{\varphi}\{\mathcal{M} \mid \mathrm{p}(\varphi) \geqslant 0\}$$
$$\cap \bigcap_{\varphi \text{ is a logical tautology}} \{\mathcal{M} \mid \mathrm{p}(\varphi) = 1\}$$
$$\cap \bigcap_{\psi \text{ and } \chi \text{ are logically incompatible}} \{\mathcal{M} \mid \mathrm{p}(\psi) + \mathrm{p}(\chi) = \mathrm{p}(\psi \vee \chi)\}$$

is an intersection of closed sets, so must be closed.

Since both $\varphi \in \mathrm{T}$ and $\varphi \notin \mathrm{T}$ are closed, and the propositional language over these is constructed with just finite operations one can prove by induction that all sentences with just say something about T are closed. Then one can show

that T being maximally consistent is closed because it can be characterised by:

$$C_{\text{max con}} = \bigcap_{\varphi} \{\mathcal{M} \mid \neg\varphi \in T \iff \varphi \notin T\}$$

$$\cap \bigcap_{\substack{\Gamma \vdash \psi \\ \Gamma \text{ is finite}}} \{\mathcal{M} \mid (\text{for each } \varphi \in \Gamma, \, \varphi \in T) \implies \psi \in T\} \qquad \square$$

The second way we strengthen the limit clause is to give a more encompassing characterisation of when C is "brought about by the sequence beneath μ". This will lead to fewer $\mathcal{M}_\mu$ satisfying the limit constraint. The definition that Belnap and Gupta focus on is that what it takes to be "brought about by the sequence beneath μ" is that the property should be *stable*. They also mention the possibility of considering the *nearly* stable properties but focus on the stable ones because that definition "is simple and general, and because it reduces to a minimum reliance on policies not absolutely dictated by the revision rule itself" (Gupta and Belnap, 1993, p. 169). We instead focus on the near stability constraint because it leads us to interesting limit-stage models, something that is very interesting but is not usually considered in much detail when working with revision theories. In their monograph Gupta and Belnap describe the distinction between stability and near stability of an "element" d, such as $T(\varphi)$, to take value $\mathbf{x}$,[8] such as false, as follows

> Stability *simpliciter* requires an element d to settle down to a value $\mathbf{x}$ after some initial fluctuations, say up to β.... In contrast, near stability allows fluctuations after β also, but these fluctuations must be confined to finite regions just after limit ordinals. (Gupta and Belnap, 1993, p. 169, their italics)

Formally the difference is described in the following definitions.

Definition 5.2.7. $C \subseteq \mathsf{Mod}$ is *stable* beneath μ in the sequence $\langle \mathcal{M}_\alpha \rangle$ if:

$$\exists \beta < \mu \, \forall \alpha \, {\substack{< \mu \\ > \beta}} \quad \mathcal{M}_\alpha \in C$$

$C \subseteq \mathsf{Mod}$ is *nearly stable* beneath μ in the sequence $\langle \mathcal{M}_\alpha \rangle$ if:

$$\exists \beta < \mu \, \forall \alpha \, {\substack{< \mu \\ > \beta}} \, \exists N_\alpha < \omega \, \forall n \, {\substack{< \omega \\ > N_\alpha}} \quad \mathcal{M}_{\alpha+n} \in C$$

If $[\varphi]$ is stable beneath μ we will say φ *is stably satisfied* beneath μ, and similarly for near stability.

An example of the difference this makes to defining the limit stages can be found in CONVERGES ALONG COPIES in Example 5.2.9. Taking the *near stability* characterisation for the limit stages will also allow us to show a desirable feature of the limit stages, namely that they satisfy Probabilistic Convention T which says that the probability of a sentence is the same as the probability that that sentence is true. Probabilistic Convention T was a motivator for Leitgeb's construction and we will discuss it in Section 5.2.2.

In conclusion, we suggest that the following definition gives a very strong but always satisfiable interpretation of the idea that $\mathcal{M}_\mu$ "sums up" the earlier improvements and we will later show that it leads to nice properties.

---

[8] Using their notation.

Whenever $C \subseteq \mathsf{Mod}$ is closed and is nearly stable beneath μ then $\mathcal{M}_\mu \in C$.

In this criterion we have asked for as many properties as possible to be taken over to limits. This is interesting because then one can determine what one can consistently ask the limit stages to be like. Given properties and behaviour that one likes, one can then consider possible weakening of the definition that still achieve such behaviour.

Our definition of a revision sequence in full is therefore the following:

Definition 5.2.8. A *revision sequence* is a sequence of models $\mathcal{M}_\alpha = (\mathrm{M}, \mathrm{T}_\alpha, \mathrm{p}_\alpha)$ such that

- $\varphi \in \mathrm{T}_{\alpha+1} \iff \mathcal{M}_\alpha \models \varphi$

- $\mathrm{p}_{\alpha+1}(\varphi)$ is the relative frequency of φ being satisfied in the (maximal) finite sequence of models leading up to $\alpha + 1$. More carefully:

$$\mathrm{p}_{\alpha+1}(\varphi) = \frac{[\![\varphi]\!]_{\mathcal{M}_{\zeta_{\alpha+1}}} + [\![\varphi]\!]_{\mathcal{M}_{\zeta_{\alpha+1}+1}} + \ldots + [\![\varphi]\!]_{\mathcal{M}_{\zeta_{\alpha+1}+k_{\alpha+1}-2}} + [\![\varphi]\!]_{\mathcal{M}_\alpha}}{k_{\alpha+1}}$$

  where $\zeta_{\alpha+1}$ is a limit ordinal and $k_{\alpha+1}$ a natural number such that $\alpha+1 = \zeta_{\alpha+1} + k_{\alpha+1}$.

- At a limit μ $\mathcal{M}_\mu$ should satisfy:

  If $C \subseteq \mathsf{Mod}$ is closed (see Definition 5.2.3) with C nearly stable in the sequence $\langle \mathcal{M}_\alpha \rangle_{\alpha < \mu}$, i.e.

  $$\exists \beta < \mu \; \forall \alpha \, {}^{<\mu}_{>\beta} \; \exists N_\alpha < \omega \; \forall n \, {}^{<\omega}_{>N_\alpha} \quad \mathcal{M}_{\alpha+n} \in C$$

  then $\mathcal{M}_\mu \in C$.

We call a sequence $\mathcal{M}_\alpha$ a *revision sequence using stability* if "nearly stable" in the above definition is replaced with "stable".

It is easy to see that a revision sequence (simpliciter, which we might instead write as "using near stability") is a revision sequence using stability.

We will note in Theorem 5.2.11 that given any M, T_0 and p_0 there will be such a revision sequence, however the revision sequence isn't uniquely determined as there will in general be infinitely many $\mathcal{M}_\mu$ to choose from at the limit stages.

To get more of a sense of this construction, we will here present some examples of how this works, particularly to illustrate the limit condition.

Example 5.2.9. Here are some examples of how the limit stage defined in Definition 5.2.8 works.[9] The only features that use the *near* stability component of the definition are Converges along copies and Non-convex.

Feature 5.2.9.1 (Fixes on a value)**.** If the probability of φ ends up fixing on some value r beneath μ, then $\mathrm{p}_\mu(\varphi) = r$.

Example. $\mathrm{p}_\omega(0=0)$. The probabilities beneath ω of $0=0$ are:

---

[9]For the examples we shall suppose that $\mathrm{T}_0 = \varnothing$ and $\mathrm{p}_0(\varphi) = 0$ for all φ
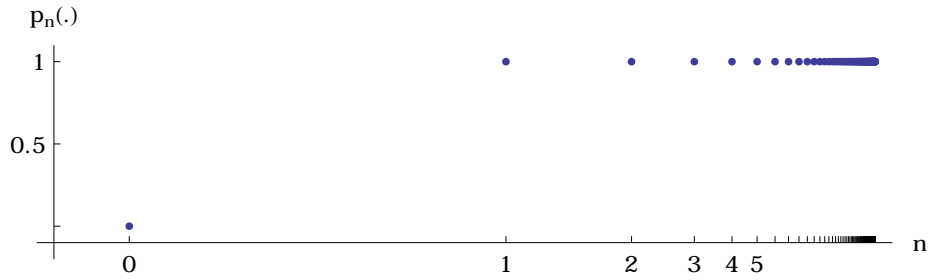
The earlier stages "bring about" the property that $p(0 = 0) = 1$ because it is stable beneath ω. It is also closed, so the limit stage $\mathcal{M}_\omega$ will also have to be in $\{\mathcal{M} \mid p(0 = 0) = 1\}$, i.e. $p_\omega(0 = 0) = 1$.

Feature 5.2.9.2 (CONVERGES TO A VALUE)**.** If the probability of φ converges to r beneath μ, then $p_\mu(\varphi) = r$.

Example. $p_\omega(\mathsf{T}\ulcorner 0 = 0\urcorner)$. The probabilities beneath ω of $\mathsf{T}\ulcorner 0 = 0\urcorner$ are:
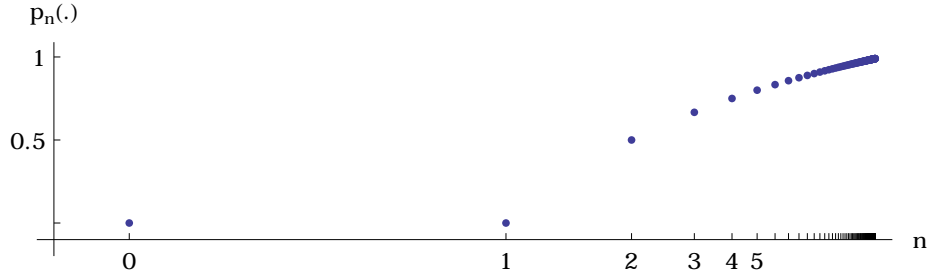
These converge to 1, and our limit constraint will lead to $p_\omega(\mathsf{T}\ulcorner 0 = 0\urcorner) = 1$. This is because for each $\epsilon > 0$ there is some N (take any $N \geqslant 1/\epsilon$) such that for all $n > N$ $p_n(\mathsf{T}\ulcorner 0 = 0\urcorner) \geqslant 1 - \epsilon$, so for each $\epsilon > 0$, $p(\mathsf{T}\ulcorner 0 = 0\urcorner) \geqslant 1 - \epsilon$ is stable beneath ω, and this is a closed property, so $\mathcal{M}_\omega \in \{\mathcal{M} \mid p(\mathsf{T}\ulcorner 0 = 0\urcorner) \geqslant 1 - \epsilon\}$ for each ϵ, i.e. $p_\omega(\mathsf{T}\ulcorner 0 = 0\urcorner) \geqslant 1 - \epsilon$, and therefore it must be that $p_\omega(\mathsf{T}\ulcorner 0 = 0\urcorner) = 1$.

Feature 5.2.9.3 (IN AN INTERVAL)**.** If the probability of φ ends up always (roughly) being in some interval, then the limit probability will also be in that interval.

Example. $p_\omega(\delta)$ with δ such that

$$\delta \leftrightarrow (\mathsf{P}\ulcorner\delta\urcorner < 0.4 \vee (0.4 \leqslant \mathsf{P}\ulcorner\delta\urcorner \leqslant 0.6 \wedge \mathsf{T}\ulcorner\delta\urcorner)).$$

Except for the first few values, the probability of δ beneath ω is always between $0.4 - \epsilon$ and $0.6 + \epsilon$ but does not converge to any value:
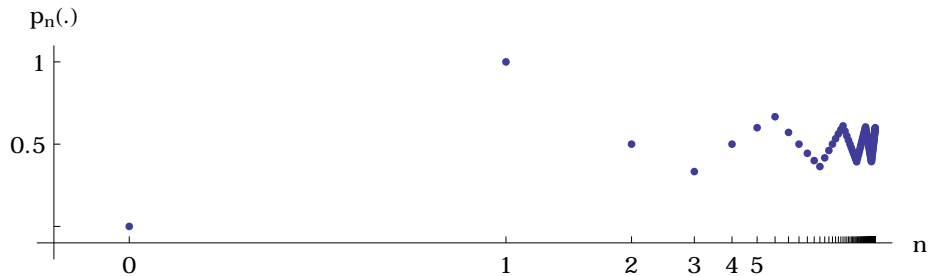
Our limit constraint requires that $0.4 \leqslant \mathrm{p}_\omega(\delta) \leqslant 0.6$ for similar reasoning as to in CONVERGES TO A VALUE.

Feature 5.2.9.4 (CONVERGES ALONG COPIES)**.** If the probability of φ converges to r along each copy of ω beneath μ, then $\mathrm{p}_\mu(\varphi) = r$.

Example. This feature is one where *near* stability plays a role. For every limit stage $\alpha + \omega$ the stability limit requirement is the same as the near stability requirement, so the difference between stability and near stability first arises at the limit stage $\omega \cdot \omega$. Consider $\mathrm{p}_{\omega \cdot \omega}(\lambda)$ with $\lambda \leftrightarrow \neg \mathsf{T}\ulcorner\lambda\urcorner$. The probabilities beneath $\omega \cdot \omega$ of λ (with $\lambda \notin \mathrm{T}_{\omega \cdot m}$ and $\mathrm{p}_{\omega \cdot m}(\lambda) = \lim_n \mathrm{p}_{\omega \cdot (m-1)+n}(\lambda) = 1/2$) are:
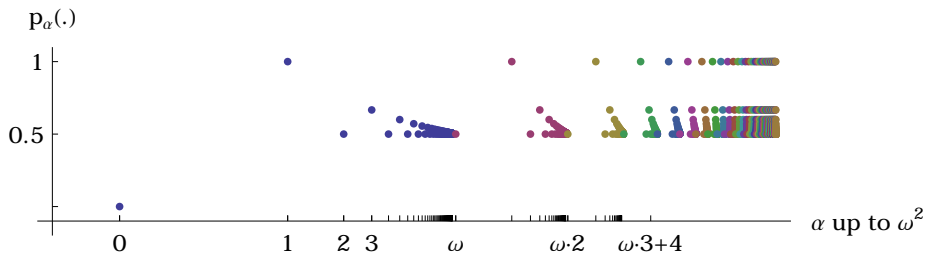
Along each copy of ω, the probability of λ converges to $1/2$. Using the near stability component of the limit definition we will get that $\mathrm{p}_{\omega \cdot \omega}(\lambda) = 1/2$. This is because each property $1/2 - \epsilon \leqslant \mathrm{p}(\lambda) \leqslant 1/2 + \epsilon$ is nearly stable beneath $\omega \cdot \omega$. More carefully, $1/2 - \epsilon \leqslant \mathrm{p}_{\alpha+n}(\lambda) \leqslant 1/2 + \epsilon$ holds whenever $n \geqslant \frac{1}{2\epsilon}$.

Feature 5.2.9.5 (RELATIONSHIPS)**.** If the probability of φ ends up always being one minus the probability of ψ beneath μ, then $\mathrm{p}_\mu(\varphi) = 1 - \mu(\psi)$.

Example. Consider $\mathrm{p}_\omega(\neg\delta)$ where δ is as in the example from IN AN INTERVAL. The probabilities of $\neg\delta$ beneath ω are:
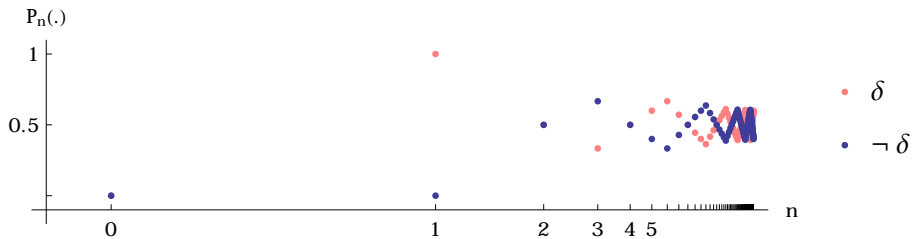
When compared to the probabilities of δ we see that for every stage beneath ω, except for the first stage, $\mathrm{p}_n(\delta) = 1 - \mathrm{p}_n(\neg\delta)$, and $\mathrm{p}(\delta) = 1 - \mathrm{p}(\neg\delta)$ is closed as mentioned in Example 5.2.6. We will therefore also have that $\mathrm{p}_\omega(\neg\delta) = 1 - \mathrm{p}_\omega(\delta)$.

Feature 5.2.9.6 (NON-CONVEX)**.**

If the probability of φ ends up always being either a or b beneath μ, then $\mathrm{p}_\mu(\varphi)$ is either a or b.

Example. Consider $p_{\omega\cdot\omega}(\delta)$ where[10]

$$\delta \leftrightarrow \left(\begin{array}{c} (\text{limstage} \wedge \neg \mathsf{T}^{\ulcorner}\delta^{\urcorner}) \\ \vee \quad (\neg\text{limstage} \wedge \mathsf{T}^{\ulcorner}\delta^{\urcorner}) \end{array} \right)$$

The probabilities of this δ beneath $\omega \cdot \omega$ are:
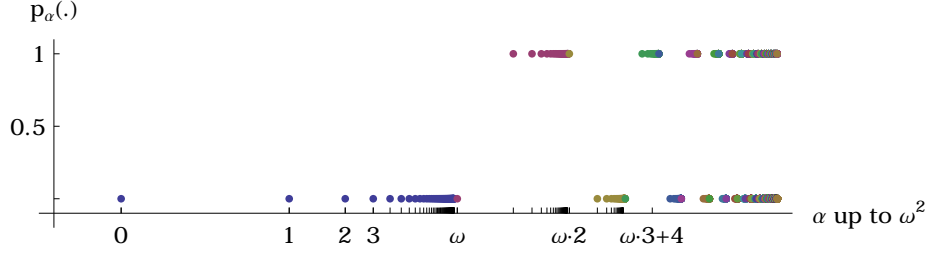
We therefore see that $p(\delta) = 0 \vee p(\delta) = 1$ is stable beneath $\omega \cdot \omega$, and this is closed as mentioned in Example 5.2.6, so we will get that $p_{\omega\cdot\omega}(\delta) = 0$ or $p_{\omega\cdot\omega}(\delta) = 1$.

We are left with one more thing to do before we can move to considering properties of the construction: we should show that this construction is satisfiable, in particular that we *can* impose such limiting behaviour.

Lemma 5.2.10. *Fix* $M_0 \in \mathsf{Mod}_{\mathcal{L}}$. *Suppose for each* $i \in I$,

$$C_i \subseteq \{(M_0, p, T) \in \mathsf{Mod} \mid \forall\varphi \in \mathsf{Sent}_{\mathsf{P},\mathsf{T}}, \; 0 \leqslant p(\varphi) \leqslant 1\},$$

and $\{C_i \mid i \in I\}$ *is a family of closed sets that has the* finite intersection property, *i.e. if for each finite* $F \subseteq I$, $\bigcap_{i\in F} C_i \neq \varnothing$.
Then $\bigcap_{i\in I} C_i \neq \varnothing$.

This says that for any M_0, $\{(M_0, p, T) \in \mathsf{Mod} \mid \forall\varphi \in \mathsf{Sent}_{\mathsf{P},\mathsf{T}}, \; 0 \leqslant p(\varphi) \leqslant 1\}$ is compact with the topology given in Definition 5.2.3.

Proof. $\{(M_0, p, T) \in \mathsf{Mod} \mid \forall\varphi \in \mathsf{Sent}_{\mathsf{P},\mathsf{T}}, \; 0 \leqslant p(\varphi) \leqslant 1\}$ is topologically equivalent to $[0,1]^{\mathsf{Sent}_{\mathsf{P},\mathsf{T}}} \times \{0,1\}^{\mathsf{Sent}_{\mathsf{P},\mathsf{T}}}$ with the product topology, which is compact by Tychonoff's theorem.[11] $\qquad\square$

Theorem 5.2.11. *Consider any* μ-*length sequence of models* $\mathcal{M}_\alpha = (M, T_\alpha, p_\alpha)$.
Then there is some $\mathcal{M}_\mu = (M, T_\mu, p_\mu)$ *such that whenever C is a closed property that is nearly stable in* $\langle\mathcal{M}_\alpha\rangle_{\alpha<\mu}$ *then* $\mathcal{M}_\mu \in C$.

Proof. Let

$$\mathcal{C} := \{C \text{ closed} \mid C \text{ is nearly stable beneath } \mu \text{ in } \langle\mathcal{M}_\alpha\rangle_{\alpha<\mu}\}$$

We will show that $\mathcal{C}$ has the finite intersection property and by Lemma 5.2.10 we will be able to deduce that

$$\bigcap \mathcal{C} \neq \varnothing,$$

---

[10] A concrete example of a formula "limstage" is $\neg\gamma$, where γ is a formula such that $\gamma \leftrightarrow \neg\forall n \mathsf{T}^{n\ulcorner}\gamma^{\urcorner}$.

[11] Which says that the product of any collection of compact topological spaces is compact with respect to the product topology.

i.e. there is some $\mathcal{M}_\mu \in \bigcap \mathcal{C}$ as required.

Let $\mathcal{D}$ be finite $\subseteq \mathcal{C}$. Enumerate the members of $\mathcal{D}$ so that $\mathcal{D} = \{C_1, \ldots, C_m\}$. We know that for each $i = 1, \ldots, m$

$$\exists \beta^i < \mu \, \forall \alpha \substack{< \mu \\ > \beta^i} \, \exists N_\alpha^i < \omega \, \forall n \substack{< \omega \\ > N_\alpha^i} \quad \mathcal{M}_{\alpha+n} \in C_i$$

Therefore for each $\alpha \geqslant \max\{\beta^i \mid i \in \{1, \ldots, m\}\}$ and $k \geqslant \max\{N_\alpha^i \mid i \in \{1, \ldots, m\}\}$

$$\mathcal{M}_{\alpha+k} \in C_1 \cap \ldots \cap C_m$$

So we can see that $\bigcap \mathcal{D} = C_1 \cap \ldots \cap C_m \neq \varnothing$. We have therefore shown that $\mathcal{C}$ has the finite intersection property, as required. $\qquad \square$

This shows that there are revision sequences as described in Definition 5.2.8.

Corollary 5.2.12. *For any* M, T_0 *and* p_0 *there is a sequence of* $\mathcal{M}_\alpha$ *with* $\mathcal{M}_0 = (M, p_0, T_0)$ *that is a revision sequence in the sense of Definition 5.2.8.*

5.2.2 Properties of the construction

The first result we will show is that in a revision sequence each p_α is a finitely additive probability and each T_α a maximally consistent set of sentences, even at the limit stages.

Theorem 5.2.13. *In any revision sequence using stability as in Definition 5.2.8, for each $\alpha > 0$, p_α is a finitely additive probability and T_α a maximally consistent set of sentences.*

Proof. We work by induction on α. For the successor stages it is easy to see that T_α is a maximally consistent set of sentences and p_α is an additive probability. For the limit stages we will show that the following criteria are closed and stably true:

1. p is a finitely additive probability.
2. T is maximally consistent.

They are stable because of the induction hypothesis.[12] They are closed as shown in Example 5.2.6. $\qquad \square$

In the traditional revision construction, the limit stages aren't maximally consistent. An advantage of having the limit stages be nice in this way is that they might themselves be considered as good candidates for an interpretation of the language. Usually the limit stages aren't taken to be suggested interpretations but are just tools for summing up the previous stages. By focusing on the limit stages themselves we can obtain some other desirable properties that aren't present at the other stages and they can therefore be seen to give at least interesting interpretations. What is interesting in this construction is that these properties of limit stages weren't hard-coded into the construction but are obtained from the more general requirement that we have given.

---

[12]In fact they are always satisfied from stage 1 on.

Although the limit probabilities will be finitely additive probabilities, we can show that they will not in general be nice with respect to the universal quantifier. In particular the interpretation of P at limits will not be what we call weak $\mathbb{N}$-additivity, and similarly the limit truth will not be ω-consistent.[13]

Definition 5.2.14. We call $\mathrm{T} \subseteq \mathsf{Sent}_{\mathsf{P},\mathsf{T}}$ ω-consistent if whenever each of

$$\varphi(\overline{0}) \in \mathrm{T},\ \varphi(\overline{1}) \in \mathrm{T},\ \varphi(\overline{2}) \in \mathrm{T} \ldots$$

then

$$\neg \forall n {\in} N \varphi(n) \notin \mathrm{T}.$$

We call $\mathrm{p} : \mathsf{Sent}_{\mathsf{P},\mathsf{T}} \to \mathbb{R}$ weakly $\mathbb{N}$-additive if whenever

$$\mathrm{p}(\varphi(\overline{0})) = 1,\ \mathrm{p}(\varphi(\overline{1})) = 1,\ \mathrm{p}(\varphi(\overline{2})) = 1,\ \ldots$$

then

$$\mathrm{p}(\neg \forall n {\in} N \varphi(n)) \neq 1.$$

Weak $\mathbb{N}$-additivity is a weak version of $\mathbb{N}$-additivity, as introduced in Definition 1.2.3. It is implied by what Leitgeb called σ-additivity in Leitgeb (2008) given the background assumption that p is finitely additive. The notion of weak $\mathbb{N}$-additivity given here could also be stated as $\{\varphi \mid \mathrm{p}(\varphi) = 1\}$ is ω-consistent.

Theorem 5.2.15. *Let $\langle \mathcal{M}_\alpha \rangle$ be a revision sequence in the sense of Definition 5.2.8. At each limit μ, p_μ will not be a weakly $\mathbb{N}$-additive probability, and each T_μ will not be ω-consistent.* [14]

This theorem and its proof can be seen as an extension of Leitgeb (2008) to the transfinite.

Proof. We prove this theorem by means of two lemmas.

Lemma 5.2.15.1. *For any sequence satisfying Definition 5.2.8,*

$$\Sigma := \{\varphi \mid \varphi \text{ is nearly stably satisfied in } \langle \mathcal{M}_\alpha \rangle_{\alpha < \mu}\}$$

is ω-inconsistent.

In fact, for γ a sentence such that

$$\gamma \leftrightarrow \neg \forall n \in N \overbrace{\mathsf{T}^{\ulcorner}\mathsf{T}^{\ulcorner} \ldots \mathsf{T}^{\ulcorner} \gamma^{\urcorner\urcorner\urcorner}}^{n+1}$$

Σ contains:

- *γ, and therefore $\neg \forall n \in N \overbrace{\mathsf{T}^{\ulcorner}\mathsf{T}^{\ulcorner} \ldots \mathsf{T}^{\ulcorner} \gamma^{\urcorner\urcorner\urcorner}}^{n+1}$*

---

[13]We would like to instead call this $\mathbb{N}$-consistency, but that would go against the tradition of the name.

[14]Equivalently, there is a sentence φ such that:

$$\mathcal{M}_\mu \models \forall n {\in} N \mathsf{T}^{\ulcorner} \varphi^{\urcorner}(n/v) \wedge \mathsf{T}^{\ulcorner} \neg \forall x {\in} N\ \varphi^{\urcorner}$$

$$\mathcal{M}_\mu \models \forall n {\in} N \mathsf{P}^{\ulcorner} \varphi^{\urcorner}(n/v) = 1 \wedge \mathsf{P}^{\ulcorner} \neg \forall x {\in} N\ \varphi^{\urcorner} = 1$$

- $\mathsf{T}\ulcorner\gamma\urcorner$
- $\mathsf{T}\ulcorner\mathsf{T}\ulcorner\gamma\urcorner\urcorner$
- ...

This is an immediate corollary of McGee (1985), a proof of which can be found in Theorem 2.4.12 (though there we used $\mathsf{P}_{=1}$ instead of T), and the result as an application to the revision theory can already be found in Gupta and Belnap (1993).

Proof. For all α, $\mathcal{M}_\alpha$ is an $\mathbb{N}$-model, therefore each $\mathcal{M}_{\alpha+1}$ satisfies:

- $\mathsf{T}\ulcorner\varphi\to\psi\urcorner\to(\mathsf{T}\ulcorner\varphi\urcorner\to\mathsf{T}\ulcorner\psi\urcorner)$
- $\mathsf{T}\ulcorner\neg\varphi\urcorner\to\neg\mathsf{T}\ulcorner\varphi\urcorner$
- $\forall n{\in}N\mathsf{T}\ulcorner\varphi\urcorner(\overline{n}/x)\to\mathsf{T}\ulcorner\forall x{\in}N\varphi\urcorner$

Therefore by the theorem in McGee (1985), which is again presented in Theorem 2.4.12, $\mathcal{M}_{\alpha+1}\models\gamma$. So by the definition of the extension of truth in a revision sequence we have that for any $n\in\mathbb{N}$, $\mathcal{M}_{\alpha+n+1}\models\mathsf{T}^n\ulcorner\gamma\urcorner$. Since this worked with arbitrary α, we have that for any α and $m>n$, $\mathcal{M}_{\alpha+m}\models\mathsf{T}^n\ulcorner\gamma\urcorner$. $\qquad\square$

Lemma 5.2.15.2. *Suppose φ is nearly stably satisfied in $\langle\mathcal{M}_\alpha\rangle_{\alpha<\mu}$ and the sequence satisfies the near stability criterion for all the closed properties. Then $\varphi\in\mathrm{T}_\mu$ and $\mathrm{p}_\mu(\varphi)=1$.*

Proof. Suppose φ is nearly stably satisfied in $\langle\mathcal{M}_\alpha\rangle_{\alpha<\mu}$. Observe that whenever φ is nearly stably satisfied then $\mathsf{T}\ulcorner\varphi\urcorner$ is also nearly stably satisfied beneath μ therefore, since $\mathsf{T}\ulcorner\varphi\urcorner$ is closed, $\varphi\in\mathrm{T}_\mu$. Similarly for each $\epsilon>0$, $1\geqslant\mathsf{P}\ulcorner\varphi\urcorner\geqslant 1-\epsilon$ is a closed and is nearly stably satisfied beneath μ, therefore for each $\epsilon>0$, $1\geqslant\mathrm{p}_\mu(\varphi)\geqslant 1-\epsilon$, i.e. $\mathrm{p}_\mu(\varphi)=1$. $\qquad\square$

Putting these lemmas together gets us

- For each $n\in\mathbb{N}$, $\mathrm{p}_\mu(\mathsf{T}^{n+1}\ulcorner\gamma\urcorner)=1$
- $\mathrm{p}_\mu(\neg\forall n{\in}N\mathsf{T}^{n+1}\ulcorner\gamma\urcorner)=1$.
- For each $n\in\mathbb{N}$, $\mathsf{T}^{n+1}\ulcorner\gamma\urcorner\in\mathrm{T}_\mu$
- $\neg\forall n\in N\,\mathsf{T}^{n+1}\ulcorner\gamma\urcorner\in\mathrm{T}_\mu$.

As required. $\qquad\square$

If we instead considered revision sequences using stability, i.e. to weaken the requirement to the stability criterion instead of the near stability criterion (still for all closed properties), one would be able to show all these results for the stages $\alpha+\omega$ but not necessarily for the other limit stages.

The last property of this construction we will discuss involves a property was very important for Leitgeb (2008, 2012), namely Probabilistic Convention T, which says that the probability of $\mathsf{T}\ulcorner\varphi\urcorner$ is the same as the probability of φ.

Definition 5.2.16. $\mathrm{p}:\mathsf{Sent}_{\mathsf{P},\mathsf{T}}\to\mathbb{R}$ satisfies *Probabilistic Convention* T if for all φ

$$\mathrm{p}(\varphi)=\mathrm{p}(\mathsf{T}\ulcorner\varphi\urcorner)$$

Probabilistic Convention T is satisfied by a model $\mathcal{M}$ of $\mathcal{L}_{\mathsf{P},\mathsf{T}}$ if $\mathcal{M}\models\mathsf{P}\ulcorner\mathsf{T}\ulcorner\varphi\urcorner\urcorner=\mathsf{P}\ulcorner\varphi\urcorner$ for all φ.

This is interesting to Leitgeb because he argues that although we cannot assign the same truth value to $\mathsf{T}\ulcorner\varphi\urcorner$ and φ, it is consistent and interesting that we can assign the same probabilities.

Theorem 5.2.17. *Probabilistic Convention T is satisfied at limit stages of revision sequences (in the sense of Definition 5.2.8), and at stages $\alpha + \omega$ for the revision sequence using stability.*

Proof. Observe that for $\epsilon > 0$, $|\mathrm{p}(\varphi) - \mathrm{p}(\mathsf{T}\ulcorner\varphi\urcorner)| \leqslant \epsilon$ is closed. We will show that it is nearly stably satisfied and therefore $|\mathrm{p}_\mu(\varphi) - \mathrm{p}_\mu(\mathsf{T}\ulcorner\varphi\urcorner)| \leqslant \epsilon$ for each $\epsilon > 0$. I.e. $\mathrm{p}_\mu(\varphi) = \mathrm{p}_\mu(\mathsf{T}\ulcorner\varphi\urcorner)$.

Each $|\mathrm{p}(\varphi) - \mathrm{p}(\mathsf{T}\ulcorner\varphi\urcorner)| \leqslant \epsilon$ is nearly stable because:

$$
\mathrm{p}_{\alpha+n}(\varphi) - \mathrm{p}_{\alpha+n}(\mathsf{T}\ulcorner\varphi\urcorner)
$$
$$
= \frac{[\![\varphi]\!]_{\mathcal{M}_{\zeta_{\alpha+n}}} + [\![\varphi]\!]_{\mathcal{M}_{\zeta_{\alpha+n+1}}} + [\![\varphi]\!]_{\mathcal{M}_{\zeta_{\alpha+n+2}}} + \ldots + [\![\varphi]\!]_{\mathcal{M}_{\alpha+n-1}}}{k_{\alpha+n}}
$$
$$
- \frac{[\![\mathsf{T}\ulcorner\varphi\urcorner]\!]_{\mathcal{M}_{\zeta_{\alpha+n}}} + [\![\mathsf{T}\ulcorner\varphi\urcorner]\!]_{\mathcal{M}_{\zeta_{\alpha+n+1}}} + \ldots + [\![\mathsf{T}\ulcorner\varphi\urcorner]\!]_{\mathcal{M}_{\alpha+n-1}}}{k_{\alpha+n}}
$$
$$
= \frac{[\![\varphi]\!]_{\mathcal{M}_{\zeta_{\alpha+n}}} + [\![\varphi]\!]_{\mathcal{M}_{\zeta_{\alpha+n+1}}} + [\![\varphi]\!]_{\mathcal{M}_{\zeta_{\alpha+n+2}}} + \ldots + [\![\varphi]\!]_{\mathcal{M}_{\alpha+n-1}}}{k_{\alpha+n}}
$$
$$
- \frac{[\![\mathsf{T}\ulcorner\varphi\urcorner]\!]_{\mathcal{M}_{\zeta_{\alpha+n}}} + [\![\varphi]\!]_{\mathcal{M}_{\zeta_{\alpha+n}}} + [\![\varphi]\!]_{\mathcal{M}_{\zeta_{\alpha+n+1}}} + \ldots + [\![\varphi]\!]_{\mathcal{M}_{\alpha+n-2}}}{k_{\alpha+n}}
$$
$$
= \frac{[\![\varphi]\!]_{\mathcal{M}_{\alpha+n-1}} - [\![\mathsf{T}\ulcorner\varphi\urcorner]\!]_{\mathcal{M}_{\zeta_{\alpha+n}}}}{k_{\alpha+n}}
$$

so $\quad |\mathrm{p}_{\alpha+n}(\varphi) - \mathrm{p}_{\alpha+n}(\mathsf{T}\ulcorner\varphi\urcorner)| \leqslant \dfrac{1}{k_\alpha + n} \longrightarrow 0 \qquad \text{as } n \longrightarrow \omega \qquad \square$

We finally mention that at each limit stage there are infinitely-many choices.

Theorem 5.2.18. *Let $\langle \mathcal{M}_\alpha \rangle_{\alpha<\mu}$ be any μ-length initial sequence of a revision sequence in the sense of Definition 5.2.8. There are infinitely many $\mathcal{M}_\mu$ which extend $\langle \mathcal{M}_\alpha \rangle_{\alpha<\mu}$ in accordance with Definition 5.2.8.*

The proof for this result can be found in Section 5.B

5.2.3 Weakening of the definition of the limit stages

In Definition 5.2.8 we presented a construction with a strong limiting criterion. However one might wish to consider possible weakenings of this definition. One reason to do this is that one disagrees with some of the behaviour which the definition leads to, alternatively one may wish to satisfy properties which are inconsistent with such a strong limiting behaviour.[15] We shall therefore here present some options for how one might weaken the limit rule.

Gupta and Belnap (1993) gives a revision theory not only for the notion of truth but for general definitions, so we can attempt to directly apply their procedure to our definition of interest. This would lead us to the stability constraint:

---

[15] We will see an example of this in Section 5.3.

If the probability of φ is stably r beneath μ then $\mathrm{p}_\mu(\varphi) = r$.

This would obtain the limiting behaviour in FIXES ON A VALUE but not even that in CONVERGES TO A VALUE, in fact almost no sentences will have their probabilities stably fix on some value. We therefore think that this is not a sufficient constraint on the limit probabilities.

To obtain the limiting behaviour in CONVERGES TO A VALUE one can instead impose the constraint:

If $\mathrm{p}_\alpha(\varphi) \longrightarrow r$ as $\alpha \longrightarrow \mu$, then $\mathrm{p}_\mu(\varphi) = r$

This would also achieve the behaviour in FIXES ON A VALUE. However it wouldn't capture any of the other behaviours so we think this is also too weak.

This also isn't phrased as a formalisation of informal limit requirement:

If a property of interest of the interpretations is "brought about" by the sequence beneath μ then it should be satisfied at the μth stage.

If we were to try to obtain this convergence behaviour as a formalisation of this informal limit requirement we would obtain:

If $r \leqslant \mathrm{p}(\varphi) \leqslant q$ is stable beneath μ, then $r \leqslant \mathrm{p}_\mu(\varphi) \leqslant q$

But this would not only obtain the behaviour of FIXES ON A VALUE and CONVERGES TO A VALUE but also that of IN AN INTERVAL. This constraint is then equivalent to:

$\liminf_{\alpha<\mu} \mathrm{p}_\alpha(\varphi) \leqslant \mathrm{p}_\mu(\varphi) \leqslant \limsup_{\alpha<\mu} \mathrm{p}_\alpha(\varphi)$

This would achieve, as special cases, the behaviours in FIXES ON A VALUE and CONVERGES TO A VALUE. One could also add to this the constraint: p_μ is a probability function. We can show that one can add the requirement that p_μ be probabilistic by using the Hahn-Banach Extension Theorem to show that there always is such an p_μ.[16] We will discuss this sort of approach in much more depth in Section 5.3.3. This kind of a proposal is a more direct generalisation of Leitgeb's construction from Leitgeb (2012). We believe that this proposal is an interesting one, it carries a significant amount of the intention of the original suggestion of stability but focuses only on the behaviour of single sentences, however we think that the behaviour in CONVERGES ALONG COPIES is desirable, so to also obtain this we instead consider *near* stability

Modifying *stability* to *near stability* also gives us the behaviour in CONVERGES ALONG COPIES. This is then:

$$\liminf_{\alpha<\mu} \liminf_{n<\omega} \mathrm{p}_{\alpha+n}(\varphi) \leqslant \mathrm{p}_\mu(\varphi) \leqslant \limsup_{\alpha<\mu} \limsup_{n<\omega} \mathrm{p}_{\alpha+n}(\varphi)$$

Which is equivalent to:

If $r \leqslant \mathrm{p}(\varphi) \leqslant q$ is nearly stable beneath μ, then $r \leqslant \mathrm{p}_\mu(\varphi) \leqslant q$

---

[16]For example see Aliprantis and Border (1999, 5.53), this can be used since $\limsup_{\alpha<\mu} x_\alpha$ is a convex function on $\mathbb{R}^\mu$.

This can also be combined with the constraint that p_μ be probabilistic, again by an application of the Hahn-Banach Extension Theorem. Choosing a p_μ which is given by a linear functional bounded above by $\limsup_{\alpha<\mu}\limsup_{n<\omega} x_{\alpha+n}$ will imply that Probabilistic Convention T holds at limit stages.[17] This is a good proposal. Although it is weaker than the definition of revision sequence we gave in Definition 5.2.8 it still leads to most of the properties we have considered.

The behaviour in RELATIONSHIPS is where we see the effect of considering the relationships between different sentences. This would not be obtained if we were only interested in properties of the form $r \leqslant p(\varphi) \leqslant q$ but is obtained because we allow the specification of a revision sequence to consider, for example, how the probability of φ relates to the probability of $\neg\varphi$. It is this feature that allows us to *derive* that the limit interpretations of P will be probabilistic and the limit interpretations of T will be a maximally consistent set of sentences. This is also what results in differences to the truth limit step, and moreover these are differences that are interesting but have not previously been suggested. This is a limit definition that I think should be further studied just in regard to its consequences for truth. The criterion seems to follow from the motivation for the usual stability clauses, namely that the limit stages should sum up properties of truth and probability that have been agreed on in the revision sequence leading up to that limit ordinal.

Finally, consider the behaviour in NON-CONVEX. It is debatable whether this behaviour is desirable. If the probabilities of φ flip between 0 and 1 then we might think that an appropriate way of summing this up would be to take the probability of φ to be 1/2. If we just focus on the probability component then we can avoid this behaviour by imposing the constraint that C be convex. More generally then we would need to say what it means to be a convex subset of Mod and impose the limit constraint only using sets which are closed and convex. We will not further consider this option in this thesis.

Although such weakenings of the limiting behaviour are interesting, there are some more fundamental worries that we have with this construction. These worries will also apply to the weakenings because the main problem they point to is the style of successor stage definition that we have used.

5.2.4 Interpretation of probability in this construction

This construction does not allow for non-trivial probabilities of contingent matters. For example consider the question of whether a coin lands heads. A sentence saying that the coin does land heads is a simple example of somewhere where we might want to assign a non-trivial probability, typically we would want to assign $p(Heads) = 1/2$. However this won't be possible in a revision sequence as we have defined it. This is because a proposition like *Heads* is part of the base language, i.e. doesn't involve reference to truth or probability, and it is therefore determined by the base model M. If $M \models Heads$ then in any revision sequence, $p_\alpha(Heads) = 1$ for each $\alpha > 0$, and if $M \models \neg Heads$ then $p_\alpha(Heads) = 0$ for all $\alpha > 0$. This shows that we can't apply these constructions to traditional uses of probabilities. The problem is that the only information we consider to determine the probability is given by the revision sequence itself, which has a background model fixed.

---

[17]Because $\limsup_{n<\omega}(p_{\alpha+n}(\mathsf{T}\ulcorner\varphi\urcorner) - p_{\alpha+n}(\varphi)) = 0$.

A related challenge is that this cannot be used to model more than one probability notion, for example to model a group of agents and their degrees of belief about one another's beliefs. This is because it is not clear how this should be extended to allow for different probability functions. The problem is that the probability functions are in a large part fixed by the base model.

We should therefore see the p that are defined in such a revision sequence as some sort of semantic probabilities, but they don't give us a theory of, for example, degrees of belief or chance.

5.2.5 Other features of the construction

Another consequence of the definitions given is that for any φ and any limit ordinal μ, either $p_{\mu+1}(\varphi) = 1$ or $p_{\mu+1}(\varphi) = 0$. It should be possible to give an alternative definition which does not have this feature while retaining the spirit of this style of revision construction, namely probabilities characterising how often something is satisfied in the revision sequence. This should be further studied.

The proof of existence of the limit stages that we gave relied on Tychonoff's theorem, which is equivalent to the axiom of choice. We therefore have not given a constructive method for picking some limit stage. In Theorem 5.2.18 we showed that there are many options to choose from when picking a limit stage. Moreover, at least in the way that we have presented it, such a non-constructive choice has to be made at *each* limit ordinal. In this sense the construction cannot be "carried out". One can instead see this in a different way: not as a *construction* of a revision sequence but instead as a *condition* for when a sequence is a revision sequence. Then we don't have to make a choice at the limits. Even with that idea, it is still interesting that there are many legitimate revision sequences, and one should still think about how they may vary.

5.3 Probabilities over possible world structures

5.3.1 Setup and successor definition

We shall now develop an alternative theory where we add some extra structure to our underlying models. This will be able to represent many different notions of probability, such as objective chances and the degrees of beliefs of agents, because we model some aspects of probability in our base models. The underlying structure will be used to give the successor definitions of probability, so it may be the case that $0 < p_{\mu+1}(\varphi) < 1$, avoiding our aforementioned worry. Our comment on the non-constructive nature of the revision definition will still apply to the revision sequences presented in this section.

Imagine two agents who have degrees of belief about something uncertain, such as whether a coin which is to be tossed will land heads or tails. They are both uncertain about this feature of the world, and they are uncertain about each other's uncertainty. Moreover, we are interested in frameworks which have some aspects of self-reference, so there are sentences such as

$$\text{Alice's degree of belief in } \pi \text{ is not great than or equal to } 1/2. \qquad (\pi)$$

In this section we will provide a revision theory that can work in such situations and can model such agents' degrees of belief.

To do this we will use the structures that we have been using throughout this thesis, namely *probabilistic modal structures*, as defined in Definition 2.1.1. These consist of a set of worlds, W, accessibility measures m_w^A, and a valuation $\mathbf{M}(w)$ that assigns to each w a model of the language without probability or truth. We shall here assume that these interpret the arithmetic vocabulary by the standard model of arithmetic and the vocabulary of the reals by the standard reals. In the alternative definition of a revision sequence that we give in this section, the probabilities of successor stages aren't given by relative frequencies, but instead we use this possible worlds framework to give meanings to P. The advantage of this construction is that it can deal with multiple interacting probability notions and it can give non-trivial probabilities to contingent vocabulary. This is because each revision sequence can now *use* the contingent vocabulary allowed in the language $\mathcal{L}$ and the revision sequence can refer to different models assigning different interpretations to this contingent vocabulary.

It can also easily be modified to result in a revision construction for modalities, like necessity or knowledge, by using an underlying Kripke structure instead of the probabilistic modal structures that we shall use.

For clarity of presentation we shall focus on structures with a single agent, or probability notion, but the definitions easily extend to multiple probability notions.

To give a revision sequence we need to define $\mathbf{p}_\alpha(w)$ and $\mathbf{T}_\alpha(w)$ for all ordinals α and worlds w. The interpretation of truth at a world has nothing to do with the other worlds, so we still define:

$$\varphi \in \mathbf{T}_{\alpha+1}(w) \iff \boldsymbol{\mathcal{M}}_\alpha(w) \models \varphi$$

where $\boldsymbol{\mathcal{M}}_\alpha(w) = (\mathbf{M}(w), \mathbf{p}_\alpha(w), \mathbf{T}_\alpha(w))$. We can easily give the definition of the successor steps by directly using the probabilistic modal structure, namely:

$$\mathbf{p}_{\alpha+1}(w)(\varphi) = m_w\{v \mid \boldsymbol{\mathcal{M}}_\alpha(v) \models \varphi\}$$

This definition is very similar to those found in, for example, Halbach et al. (2003) where they gave a similar definition for the case of necessity.

For example this will result in a revision sequence as in Fig. 5.1.

We now have to explain how to give the limit stages.

One could take some alternative way of defining limit truth, for example as in the standard revision sequences, and then define a limit probability derivative on that, by letting

$$\mathbf{p}_\mu(w)(\varphi) = m_w\{v \mid \varphi \in \mathbf{T}_\mu(v)\}$$

Such a construction will allow a nice connection between the interpretation of truth and probability at the limit stage but just focusing on probability it has some odd consequences. For example, it might be the case that $\mathbf{p}_\omega(w)(\lambda) = 0$ even though for each $n < \omega$, $\mathbf{p}_n(w)(\lambda) = 1/2$.[18]

To ensure that such situations do not arise, we will instead focus on limit stage proposals that work in the spirit of the limit stage that we defined for Section 5.2.
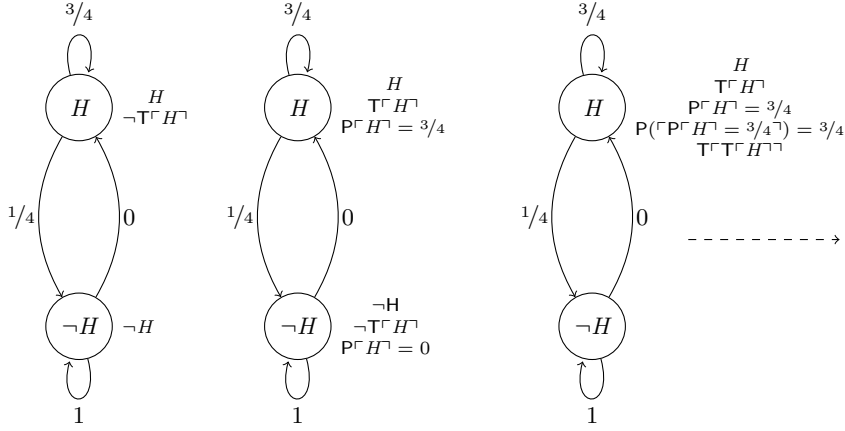
---

[18] For example:

Figure 5.1: Example of a revision sequence with a probabilistic modal structure

## 5.3.2 Limit stages "sum up" previous stages

We might therefore instead propose an alternative limit definition for probability by directly applying our limit stage construction from Definition 5.2.8.

> If $C \subseteq \mathsf{Mod}$ is closed with C nearly stable in the sequence $\langle \boldsymbol{\mathcal{M}}_\alpha(w) \rangle_{\alpha < \mu}$, then $\boldsymbol{\mathcal{M}}_\mu(w) \in C$.

This definition is still possible since the proof in Theorem 5.2.11 will still apply. Similarly, Theorem 5.2.13 will still hold so the limit probabilities will be finitely additive probability functions.

For certain matters this may seem too weak: Given the additional structure there are additional properties of the sequence of previous interpretations that might be "brought about by the previous stages" but not yet required to be satisfied at the limit stage. Such properties will relate to the relations between truth and probabilities at the different worlds. For example consider the situation described in Fig. 5.2 with $\lambda \in \mathbf{T}_0(w_{Circle})$ and $\lambda \notin \mathbf{T}_0(w_{Cross})$.

Consider $\mathbf{p}_\omega(w)(\lambda)$. Beneath ω, $\mathbf{p}_n(w)(\lambda)$ how it acts is described in Fig. 5.3.
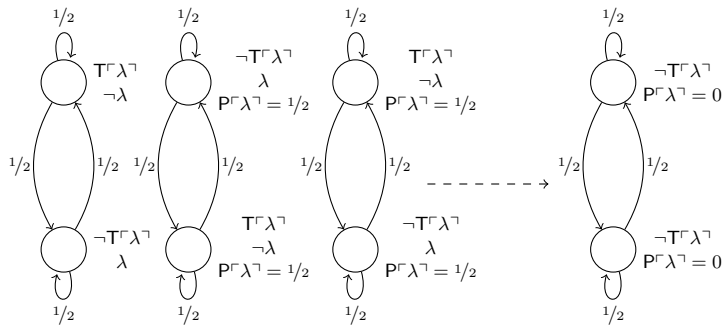
This example relies on taking $\mathbf{T}_\mu$ to be the collection of sentences that are nearly stably true beneath μ, though this undesirable feature shouldn't rely on this definition of limit-truth.

139

Figure 5.2: The probabilistic modal structure showing that some properties aren't taken to limits.
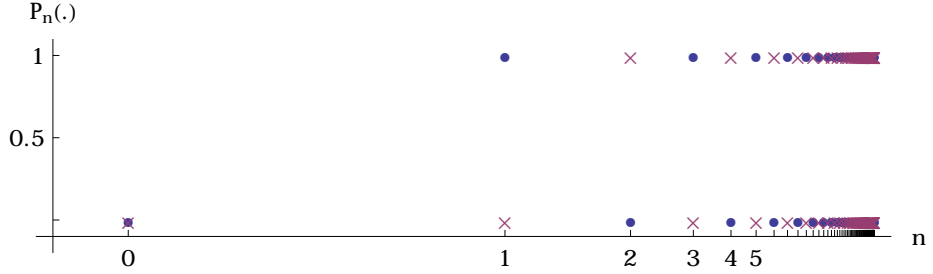
Figure 5.3: Revising probability showing that some properties aren't taken to limits.

So a property of the probabilities which is brought about by the sequence beneath ω is that the probability of λ in w_{Circle} is one minus the probability of λ in w_{Cross}, i.e. $\mathrm{p}(w_{Circle})(\lambda) = 1 - \mathrm{p}(w_{Cross})(\lambda)$. We might therefore ask that this is also satisfied at the limit stage.

To take account of this we can extend the limit constraint. To do this, though, we first need to extend the notion of being closed from Definition 5.2.3. This definition will have that **Mod** is topologically equivalent to Mod, which is in turn equivalent to $\mathsf{Mod}_{\mathcal{L}}^{W} \times \mathbb{R}^{\mathsf{Sent}_{\mathsf{P},\mathsf{T}} \times W} \times \{0, 1\}^{\mathsf{Sent}_{\mathsf{P},\mathsf{T}} \times W}$.

Definition 5.3.1. Let **Mod** be all functions assigning to each $w \in W$ some member of Mod. So for $\boldsymbol{\mathcal{M}} \in \mathbf{Mod}$, $\boldsymbol{\mathcal{M}}(w) = (\mathbf{M}(w), \mathbf{p}(w), \mathbf{T}(w)) \in \mathsf{Mod}$.

We say that $\mathbf{C} \subseteq \mathbf{Mod}$ is closed iff it is closed in the product topology on **Mod**. This can be defined by: $\mathbf{C} \subseteq \mathbf{Mod}$ *is closed* if

- There is some C a closed subset of Mod, $\mathbf{C} = \{\boldsymbol{\mathcal{M}} \mid \boldsymbol{\mathcal{M}}(w) \in C\}$.[19]

- $\mathbf{C} = \mathbf{A} \cup \mathbf{B}$ for some $\mathbf{A}, \mathbf{B}$ closed,

- $\mathbf{C} = \bigcap_{i \in I} \mathbf{A}_i$ with each $\mathbf{A}_i$ closed.

We can then extend the limit stage definition by:

If $\mathbf{C} \subseteq \mathbf{Mod}$ is closed with $\mathbf{C}$ nearly stable in the sequence $\langle \boldsymbol{\mathcal{M}}_\alpha \rangle_{\alpha < \mu}$, then $\boldsymbol{\mathcal{M}}_\mu \in \mathbf{C}$.

We then result in the following definition of a revision sequence:

---

[19]This could equivalently be defined by: $\mathbf{C}$ takes the form:

- $\mathbf{M}(w_0) \neq \mathrm{M}_0$,

- $\varphi \in \mathbf{T}(w)$, or $\varphi \notin \mathbf{T}(w)$, or

- $r \leqslant \mathbf{p}(w)(\varphi)$, or $r \geqslant \mathbf{p}(w)(\varphi)$,

where, as before, these describe sets, e.g. $\varphi \in \mathbf{T}(w)$ describes the set $\{\boldsymbol{\mathcal{M}} \in \mathbf{Mod} \mid \varphi \in \mathbf{T}(w)\}$.

Definition 5.3.2. A *revision sequence over a probabilistic modal structure* $\mathfrak{M}$ is a sequence of models $\boldsymbol{\mathcal{M}}_\alpha \in \mathbf{Mod}$, $\boldsymbol{\mathcal{M}}_\alpha(w) = (\mathbf{M}(w), \mathbf{p}_\alpha(w), \mathbf{T}_\alpha(w))$ such that

- $\varphi \in \mathbf{T}_{\alpha+1}(w) \iff \boldsymbol{\mathcal{M}}_\alpha(w) \models \varphi$

- $\mathbf{p}_{\alpha+1}(w)(\varphi) = m_w\{v \mid \boldsymbol{\mathcal{M}}_\alpha(v) \models \varphi\}$

- At a limit μ, $\boldsymbol{\mathcal{M}}_\mu$ should satisfy:

  If $\mathbf{C} \subseteq \mathbf{Mod}$ is closed (see Definition 5.3.1) with $\mathbf{C}$ nearly stable in the sequence $\langle \boldsymbol{\mathcal{M}}_\alpha \rangle_{\alpha < \mu}$, i.e.

  $$\exists \beta < \mu \, \forall \alpha \mathop{>}\limits_{\beta}^{<\mu} \exists N_\alpha < \omega \, \forall n \mathop{>}\limits_{N_\alpha}^{<\omega} \quad \boldsymbol{\mathcal{M}}_{\alpha+n} \in \mathbf{C}$$

  then $\boldsymbol{\mathcal{M}}_\mu \in \mathbf{C}$.

This will be satisfiable because the proof in Theorem 5.2.11 will still apply. Moreover, we will still have all the results Theorems 5.2.13, 5.2.15 and 5.2.18 as the proofs with go through in this more general setting.[20] However, we will *not* have Theorem 5.2.17, i.e. some of the revision sequences will fail to satisfy Probabilistic Convention T. For someone who is worried about this, we can weaken the constraint so as to have Probabilistic Contention T satisfied. A result of that modification, though, is that one won't have that the limit stages must be probabilistic and maximally consistent automatically following from the definition, they instead will have to be built in. It is this option which we study in the next section.

5.3.3 Limit stages summing up – a weaker proposal using Banach limits so we can get Probabilistic Convention T.

In the previous section we considered a strengthening of the criterion:

  If $C \subseteq \mathsf{Mod}$ is closed with C nearly stable in the sequence $\langle \boldsymbol{\mathcal{M}}_\alpha(w) \rangle_{\alpha < \mu}$, then $\boldsymbol{\mathcal{M}}_\mu(w) \in C$.

This leads to a failure of Probabilistic Convention T. In fact the criterion, even un-strengthened, leads to failures of Probabilistic Convention T, i.e. for some φ $\mathbf{p}_\mu(w)(\varphi) \neq \mathbf{p}_\mu(w)(\mathsf{T}\ulcorner\varphi\urcorner)$. Theorem 5.2.17 fails in this framework because now $|\mathbf{p}(w)(\varphi) - \mathbf{p}(w)(\mathsf{T}\ulcorner\varphi\urcorner)| \leqslant \epsilon$ is not (nearly) stable beneath limits in $\langle \boldsymbol{\mathcal{M}}_\alpha(w) \rangle_{\alpha < \mu}$ because our definition of the successor stages has changed. We can see that in fact we get bad failures of Probabilistic Convention T.

Example 5.3.3. For example consider $\mathfrak{M}_{\mathsf{omn}}$ containing one world w_0, and observe that after stage 2, $\mathbf{p}_n(w_0)(\lambda) = 1 - \mathbf{p}_n(w_0)(\mathsf{T}\ulcorner\lambda\urcorner)$ as can be seen in Fig. 5.4.

Moreover $\mathbf{p}(w_0)(\lambda) = 1 - \mathbf{p}(w_0)(\mathsf{T}\ulcorner\lambda\urcorner)$ is a closed property that is stable in this $\langle \boldsymbol{\mathcal{M}}_n(w_0) \rangle_{n < \omega}$ and therefore our constraint would *require* that $\mathbf{p}_\omega(w_0)(\lambda) = 1 - \mathbf{p}_\omega(w_0)(\mathsf{T}\ulcorner\lambda\urcorner)$. This will also be the case at any limit ordinal, showing that this criterion requires bad failures of Probabilistic Convention T.

---

[20]For the proof of Theorem 5.2.15, one needs to redefine Σ as $\{\varphi \mid \varphi$ is *uniformly* nearly stably satisfied in $\langle \mathbf{M}_\alpha \rangle\}$, where we say φ is *uniformly* nearly stably satisfied if $\{\mathbf{M} \mid$ for all w, $\mathbf{M}(w) \models \varphi\}$ is nearly stable.
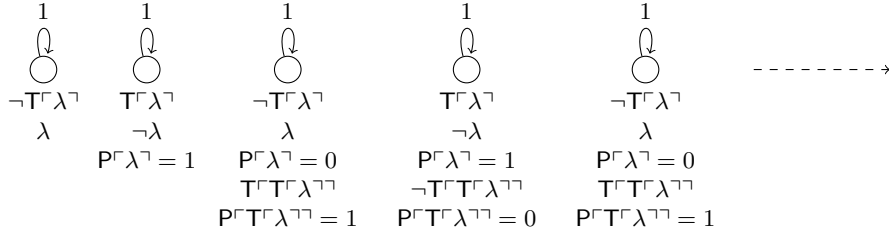
Figure 5.4: Revising the probability of the liar in $\mathfrak{M}_{\mathsf{omn}}$.

If one is interested in keeping Probabilistic Convention T, there is a way of weakening our constraint that would allow for Probabilistic Convention T to be satisfiable. This would be to instead just focus on single sentences and impose the constraint that

$$\liminf\nolimits_{\alpha<\mu} \mathbf{p}_\alpha(w)(\varphi) \leqslant \mathbf{p}_\mu(w)(\varphi) \leqslant \limsup\nolimits_{\alpha<\mu} \mathbf{p}_\alpha(w)(\varphi)$$

or equivalently:

If $r \leqslant \mathrm{p}(\varphi) \leqslant q$ is stable in $\langle \boldsymbol{\mathcal{M}}_\alpha(w)\rangle_{\alpha<\mu}$, then it should be satisfied at μ.

Imposing this constraint will lead to the behaviour in FIXES ON A VALUE, CONVERGES TO A VALUE and IN AN INTERVAL. Modifying it to understand it with *near* stability will also lead to CONVERGES ALONG COPIES.

We will be able to show that we can always find limit stages which:

- Have $\mathbf{p}_\mu$ probabilistic,

- Satisfy Probabilistic Convention T, i.e. $\mathbf{p}_\mu(\varphi) = \mathbf{p}_\mu(\mathsf{T}\ulcorner\varphi\urcorner)$,

- Satisfy the limit constraint:

    If $r \leqslant \mathrm{p}(\varphi) \leqslant q$ is nearly stable in $\langle \boldsymbol{\mathcal{M}}_\alpha(w)\rangle_{\alpha<\mu}$, then it should be satisfied at μ.

We do this by choosing limit stages by means of so called *Banach limits*. We will take:
$$\mathbf{p}_\mu(w)(\varphi) = \mathsf{BanLim}_\mu\left(\langle\mathbf{p}_\alpha(w)(\varphi)\rangle_{\alpha<\mu}\right)$$

where BanLim_μ is some functional which is linear, positive, normalised, finitely shift-invariant and satisfies near stability. This could be proposed as an appropriate generalisation of the notion of a Banach limit to the transfinite.

Theorem 5.3.4. *For every limit ordinal μ we can find some BanLim_μ defined on the space of bounded μ-length sequences of real numbers which is:*

- *linear, i.e.*

$$\mathsf{BanLim}_\mu\left(\langle rx_\alpha + qy_\alpha\rangle_{\alpha<\mu}\right) = r{\cdot}\mathsf{BanLim}_\mu\left(\langle x_\alpha\rangle_{\alpha<\mu}\right) + q{\cdot}\mathsf{BanLim}_\mu\left(\langle y_\alpha\rangle_{\alpha<\mu}\right)$$

- *positive, i.e. if each $x_\alpha \geqslant 0$ then $\mathsf{BanLim}_\mu\left(\langle x_\alpha\rangle_{\alpha<\mu}\right) \geqslant 0$*

142

- *normalized, i.e.* $\mathsf{BanLim}_\mu\left(\mathbf{1}\right) = 1$

- *(finitely) shift-invariant, i.e.*

$$\mathsf{BanLim}_\mu\left(\langle x_\alpha\rangle_{\alpha<\mu}\right) = \mathsf{BanLim}_\mu\left(\langle x_{\alpha+1}\rangle_{\alpha<\mu}\right)$$

- *satisfies near stability, i.e.*

$$\liminf_{\alpha<\mu}\liminf_{n<\omega} x_{\alpha+n} \leqslant \mathsf{BanLim}_\mu\left(\langle x_\alpha\rangle_{\alpha<\mu}\right) \leqslant \limsup_{\alpha<\mu}\limsup_{n<\omega} x_{\alpha+n}$$

The linear positive, normalised components of the definition of a Banach limit will get us that $\mathbf{p}_\mu$ is probabilistic, the shift-invariance will get us Probabilistic Convention T, and the near stability component will get us the restricted version of the near stability limit requirement where we only consider properties of the form $r \leqslant \mathrm{p}(\varphi) \leqslant q$. So this theorem will show us that we can choose limit stages with the desired properties.

To show the existence of such a Banach limit we generalise one of the usual proofs of the existence of Banach limits. When we considered revision sequences as defined in Definition 5.2.8, we obtained Probabilistic Convention T at limit stages. In the proof of this, Theorem 5.2.17, we only used the near stability requirement for properties of the form $|\mathrm{p}(\varphi) - \mathrm{p}(\mathsf{T}^\ulcorner\varphi^\urcorner)| \leqslant \epsilon$, or, equivalently $-\epsilon \leqslant \mathsf{P}^\ulcorner\varphi^\urcorner - \mathsf{P}^\ulcorner\mathsf{T}^\ulcorner\varphi^\urcorner^\urcorner \leqslant \epsilon$, and to see that each of these is nearly stable we use the fact that the successor stages are defined as relative frequencies and that $[\![\varphi]\!]_{\mathcal{M}_\alpha} = [\![\mathsf{T}^\ulcorner\varphi^\urcorner]\!]_{\mathcal{M}_{\alpha+1}}$. When we consider the weakening of revision sequences which only uses the near stability requirement for properties of the form $r \leqslant \mathrm{p}(\varphi) \leqslant q$, one can see $\mathrm{p}_\mu(\varphi)$ as a function of $\langle[\![\varphi]\!]_{\mathcal{M}_\alpha}\rangle_{\alpha<\mu}$. In that context Probabilistic Convention T expressed a finitely shift-invariant property. So obtaining limit probabilities using the near stability criterion for properties of the form $r \leqslant \mathrm{p}(\varphi) \leqslant q$ was very close to showing the existence of a Banach limit. To formalise that we will mimic the definition of such a revision sequence.

Proof. Let l^μ denote the space of bounded μ-length sequences of real numbers. Define $f : l^\mu \to \mathbb{R}$ by[21]

$$f\left(\langle x_\alpha\rangle_{\alpha<\mu}\right) = \limsup_{\mu'<\mu}\limsup_{n<\omega} \mathsf{Av}_{\mu'+n}\left(\langle x_\beta\rangle_{\beta<\mu'+n}\right)$$

where

$$\mathsf{Av}_{\mu'+n}\left(\langle x_\beta\rangle_{\beta<\mu'+n}\right) := \frac{x_{\mu'} + \ldots + x_{\mu'+n-1}}{n}$$

f is a sub-linear functional since $\limsup$ and Av_α are sub-linear functionals, and therefore the composition of them is also a sub-linear functional.

Therefore by the Hahn-Banach extension theorem[22] there is some BanLim_μ a linear functional defined on l^μ which is dominated by f. One can check

---

[21] We use this more complicated definition of f instead of the more obvious definition as $\limsup_{\alpha<\mu} \mathsf{Av}_\alpha\left(\langle x_\beta\rangle_{\beta<\alpha}\right)$ which would be a direct generalisation of the proof for BanLim_ω since the latter definition is only shift invariant at ordinals $\alpha + \omega$.

[22] For example, see Aliprantis and Border (1999, 5.53).

that this BanLim_μ has the required properties. For example to show that it is finitely shift-invariant:

$$\left| \mathsf{Av}_{\alpha+n} \left(\langle x_\beta - x_{\beta+1} \rangle_{\beta < \alpha+n} \right) \right|$$
$$= \frac{(x_{\zeta_\alpha} - x_{\zeta_\alpha+1}) + \ldots + (x_{\alpha+n} - x_{\alpha+n+1})}{k_{\alpha+n}}$$
$$= \frac{x_{\zeta_\alpha} - x_{\alpha+n+1}}{k_{\alpha+n}}$$
$$\longrightarrow 0 \text{ as } n \longrightarrow \omega \text{ since } x_\alpha \text{ is bounded}$$

So

$$\mathsf{BanLim}_\mu(\langle x_\alpha \rangle_{\alpha<\mu}) - \mathsf{BanLim}_\mu(\langle x_{\alpha+1} \rangle_{\alpha<\mu}) = \mathsf{BanLim}_\mu(\langle x_\alpha - x_{\alpha+1} \rangle_{\alpha<\mu})$$
$$\leqslant f\left(\langle x_\alpha - x_{\alpha+1} \rangle_{\alpha<\mu} \right) = 0$$

as required. $\qquad\square$

We have now presented two different options for how to define limit probabilities using the limiting behaviour of the sequence. The first suggestion, in Section 5.3.2, was a criterion which forces strong limiting clause given the behaviour of probability in the preceding sequence. This was given by:

> If $\mathbf{C} \subseteq \mathbf{Mod}$ is closed with $\mathbf{C}$ nearly stable in the sequence $\langle \boldsymbol{\mathcal{M}}_\alpha \rangle_{\alpha<\mu}$, then $\boldsymbol{\mathcal{M}}_\mu \in \mathbf{C}$.

Our second suggestion weakened the limiting behaviour. For some closed properties which are stably satisfied beneath μ were not asked to be satisfied at the limit stages. For example $\mathrm{p}(\mathsf{T}^\ulcorner\lambda^\urcorner) = 1 - \mathrm{p}(\lambda)$ might be stably satisfied beneath μ, but if we want the limit stages to satisfy Probabilistic Convention T we will require that this is *not* satisfied at the limit stages as we instead want $\mathrm{p}(\mathsf{T}^\ulcorner\lambda^\urcorner) = \mathrm{p}(\lambda)$ to be satisfied. So we dropped some of the limiting behaviour and only focused on properties of the form $r \leqslant \mathrm{p}(\varphi) \leqslant q$ and instead imposed an alternative additional constraint: that the limit probabilities satisfy Probabilistic Convention T and be probabilistic. We showed that it is possible to require such limiting behaviour by generalising the notion of a Banach limit to arbitrary length sequences. It would be interesting to see what further limiting behaviour is consistent with Probabilistic Convention T.

I find Probabilistic Convention T a nice feature, but not enough to take this alternative approach to the limit stages, unless of course one has specific reasons to be interested in Probabilistic Convention T. The definition which I find the most attractive is the one presented in Section 5.3.2 given by considering the nearly stable closed properties. A particularly nice feature of this definition is that the criterion is just a single constraint (unlike the version to get Probabilistic Convention T where one has to explicitly require an extra feature) and has many desirable consequences, for example that the limit stages are finitely additive probabilities and finitely consistent extensions of truth.

5.4 Theories for these constructions

5.4.1 In the general case

It is useful to give theories explaining properties of these revision sequences. This is something that is considered when one works with revision theories for truth, and we do the same here for revision theories of probability. In Leitgeb (2012) Leitgeb gives some axiomatic theories which directly apply to our notions of a revision sequence.

Theorem 5.4.1. *In any revision sequence in the sense of Definition 5.2.8, the theorems of the axiomatic theory PT_1 from Leitgeb (2012) are nearly stably satisfied in any sequence up to any limit ordinal I.e. if $\mathsf{PT}_1 \vdash \varphi$ then*

$$\exists \beta < \mu \; \forall \alpha_{>\beta}^{<\mu} \; \exists N < \omega \; \forall n_{>N}^{<\omega} \quad \mathcal{M}_{\alpha+n} \models \varphi$$

PT_1 *is:*

- *Base theory:*

    - $\mathsf{PA}^{\mathcal{L}_{\mathsf{P},\mathsf{T}}}$, *Peano Arithmetic with the induction axioms extended to the full language of $\mathcal{L}_{\mathsf{P},\mathsf{T}}$,*[23]
    - *The theory of real closed fields,* ROCF.

- *Axioms and rules for truth:*

    - *The commutation axioms for T with respect to $\neg, \wedge, \vee, \forall, \exists$,*
    - *All T-biconditionals for atomic sentences which do not involve T or P,*
    - $\dfrac{\varphi}{\mathsf{T}\ulcorner\varphi\urcorner}$ *and* $\dfrac{\mathsf{T}\ulcorner\varphi\urcorner}{\varphi}$.

- *Axioms and rules for each probability function symbol P:* [24]

    - $\forall\varphi(0 \leqslant \mathsf{P}\ulcorner\varphi\urcorner \leqslant \overline{1})$,
    - $\forall\varphi(\mathsf{Prov}_{\mathsf{PA}^{\mathcal{L}_{\mathsf{P},\mathsf{T}}}}\ulcorner\varphi\urcorner \to \mathsf{P}\ulcorner\varphi\urcorner = \overline{1})$,
    - $\forall\varphi\forall\psi(\mathsf{Prov}_{\mathsf{PA}^{\mathcal{L}_{\mathsf{P},\mathsf{T}}}}\ulcorner\neg(\varphi \wedge \psi)\urcorner \to \mathsf{P}\ulcorner\varphi\urcorner + \mathsf{P}\ulcorner\psi\urcorner = \mathsf{P}\ulcorner\varphi \vee \psi\urcorner)$,
    - $\forall\varphi(\mathsf{P}\ulcorner\forall v \in N(\varphi(v))\urcorner = \lim_{n\longrightarrow\infty} \mathsf{P}\ulcorner\varphi(\overline{1}) \vee \ldots \vee \varphi(\overline{n})\urcorner)$,[25]
    - $\dfrac{\varphi}{\mathsf{P}\ulcorner\varphi\urcorner > 1 - 1/\overline{n}}$,
    - $\dfrac{\mathsf{P}\ulcorner\varphi\urcorner = 1}{\varphi}$.

    - *Approximation to Probabilistic Convention T*

---

[23]Quantification in the theory of Peano Arithmetic is understood as quantification restricted by the natural number predicate N.

[24] "$\forall\varphi$" is understood as $\forall x(\mathsf{Sent}_{\mathsf{P},\mathsf{T}}(x) \to \ldots)$. Quantification into $\ulcorner.\urcorner$ can be made precise using standard arithmetic means. For further details of a way to do this see Halbach (2011). We are following Leitgeb (2012) in using this notation. For example, $\forall\varphi(0 \leqslant \mathsf{P}\ulcorner\varphi\urcorner \leqslant 1)$ is shorthand for $\forall x(\mathsf{Sent}_{\mathsf{P},\mathsf{T}}(x) \to (0 \leqslant \mathsf{P}x \leqslant 1)$ and $\forall\varphi(|\mathsf{P}\ulcorner\mathsf{T}\ulcorner\varphi\urcorner\urcorner - \mathsf{P}\ulcorner\varphi\urcorner| < 1/\overline{n})$ would appropriately be formulated by: $\forall x(\mathsf{cClTerm}_{L_{\mathsf{PA}}}(x) \to |\mathsf{PT}x - \mathsf{P}x^\circ| < 1/\overline{n})$

[25]lim can be made precise using a usual $\epsilon - \delta$ definition of convergence.

$$- \forall\varphi(|\mathsf{P}^\ulcorner\mathsf{T}^\ulcorner\varphi^\urcorner\urcorner - \mathsf{P}^\ulcorner\varphi^\urcorner| < 1/\overline{n})$$

If we just focus on limit stages we can get a stronger theory being satisfied.

Theorem 5.4.2. *In any revision sequence in the sense of Definition 5.2.8[26], the theorems of the axiomatic theory* PT'_2 *is satisfied at every limit ordinal.*
PT'_2 *is:*

- *Base theory as before.*

- *Axioms and rules for truth:*

    - *The commutation axioms for* T *with respect to* $\neg, \wedge, \vee$,[27]
    - *All* T-*biconditionals for atomic sentences which do not involve* T *or* P
    - $\forall\varphi(\mathsf{Prov}_{\mathsf{PT}_1}{}^\ulcorner\varphi^\urcorner \to \mathsf{T}^\ulcorner\varphi^\urcorner).$

- *Axioms for* P*:* [28]

    - $\forall\varphi(0 \leqslant \mathsf{P}^\ulcorner\varphi^\urcorner \leqslant 1),$
    - $\forall\varphi(\mathsf{Prov}_{\mathsf{PT}_1}{}^\ulcorner\varphi^\urcorner \to \mathsf{P}^\ulcorner\varphi^\urcorner = 1),$
    - $\forall\varphi\forall\psi(\mathsf{Prov}_{\mathsf{PT}_1}{}^\ulcorner\neg(\varphi \wedge \psi)^\urcorner \to \mathsf{P}^\ulcorner\varphi^\urcorner + \mathsf{P}^\ulcorner\psi^\urcorner = \mathsf{P}^\ulcorner\varphi \vee \psi^\urcorner),$
    - $\forall\varphi\forall a, b \in R(\mathsf{Prov}_{\mathsf{PT}_1}{}^\ulcorner r \leqslant \mathsf{P}^\ulcorner\varphi^\urcorner \leqslant q^\urcorner \to r \leqslant \mathsf{P}^\ulcorner\varphi^\urcorner \leqslant q).$

- *Probabilistic Convention* T*:*

    - $\forall\varphi(\mathsf{P}^\ulcorner\mathsf{T}^\ulcorner\varphi^\urcorner\urcorner = \mathsf{P}^\ulcorner\varphi^\urcorner).$

- *A version of Miller's Principle:*

    - $\forall\varphi(\mathsf{P}^\ulcorner\mathsf{P}^\ulcorner\varphi^\urcorner = 1^\urcorner > 0 \to \mathsf{P}({}^\ulcorner\mathsf{T}^\ulcorner\varphi^\urcorner\urcorner \,|\, {}^\ulcorner\mathsf{P}^\ulcorner\varphi^\urcorner = 1^\urcorner)) = 1.$

This extends Leitgeb's PT_2 by also including axioms and rules for truth. If near stability is replaced with stability, PT_2 will be satisfied at limit ordinals of the form $\alpha + \omega$.

For the probabilistic modal structure construction we can make the following modifications:

Theorem 5.4.3. *Fix any probabilistic modal structure. In a revision sequence using as successor rules:*

$$\varphi \in \mathbf{T}_{\alpha+1}(w) \iff \boldsymbol{\mathcal{M}}_\alpha(w) \models \varphi$$

$$\mathbf{p}_{\alpha+1}(w)(\varphi) = m_w\{v \,|\, \boldsymbol{\mathcal{M}}_\alpha(v) \models \varphi\}$$

the theorems of the axiomatic theory $\mathsf{PT}_1^{\mathsf{PMS}}$ *are uniformly* *nearly stably satisfied in any sequence up to any limit ordinal. I.e. if* $\mathsf{PT}_1 \vdash \varphi$ *then*

$$\exists\beta < \mu \,\forall\alpha_{>\beta}^{<\mu} \,\exists N < \omega \,\forall n_{>N}^{<\omega} \,\forall w \in W \quad \boldsymbol{\mathcal{M}}_{\alpha+n}(w) \models \varphi$$

$\mathsf{PT}_1^{\mathsf{PMS}}$ *modifies* PT_1 *by:*

---

[26]The proofs only rely on having the near stability constraint for properties of the form $r \leqslant \mathsf{P}^\ulcorner\varphi^\urcorner \leqslant q$ and that the p_μ is a finitely additive probability, and T_μ is a maximally consistent set of sentences.

[27]but not for $\forall$ or $\exists$

[28] "$\forall\varphi$" is really understood as quantification over natural numbers that are codes of sentences of $\mathcal{L}_{\mathsf{P,T}}$. Quantification into $^\ulcorner.^\urcorner$ can be made precise using standard arithmetic means.

- *Base theory as in* PT_1,

- *Axioms and rules for truth as in* PT_1,

- *Axioms and rules for each probability function symbol* P: [29] *Modifies those from* PT_1 *by now using the alternative rules*

  - $$\frac{\varphi}{\mathsf{P}\ulcorner\varphi\urcorner = 1}$$

  - *If the fixed probabilistic modal structure is such that for all w there is some v with $m_v^{\mathrm{A}}\{w\} > 0$, then we also have:* $\dfrac{\mathsf{P}\ulcorner\varphi\urcorner = 1}{\varphi}$

- *Modified version of Probabilistic Convention* T

  - $\forall\varphi\forall a \in R(\mathsf{T}\ulcorner\mathsf{P}\ulcorner\varphi\urcorner = r\urcorner \leftrightarrow \mathsf{P}\ulcorner\mathsf{T}\ulcorner\varphi\urcorner\urcorner = r)$

Here we have replaced the axiom $\forall\varphi(|\mathsf{P}\ulcorner\mathsf{T}\ulcorner\varphi\urcorner\urcorner - \mathsf{P}\ulcorner\varphi\urcorner| < 1/\overline{n})$ by $\forall\varphi\forall a \in R(\mathsf{T}\ulcorner\mathsf{P}\ulcorner\varphi\urcorner = r\urcorner \leftrightarrow \mathsf{P}\ulcorner\mathsf{T}\ulcorner\varphi\urcorner\urcorner = r)$. This is in some sense stronger and in another sense weaker. It is stronger because the approximate equality is replaced by an exact equality, but weaker because we also need $\mathsf{P}\ulcorner\varphi\urcorner$ to appear inside of a truth predicate instead of just being able to state $\mathsf{P}\ulcorner\varphi\urcorner = \mathsf{P}\ulcorner\mathsf{T}\ulcorner\varphi\urcorner\urcorner$.

We can also consider the theory at the limit stages, which has some nice features:

Theorem 5.4.4. *If we take a revision sequence based on a probabilistic modal structure with the successor stages*

$$\varphi \in \mathbf{T}_{\alpha+1}(w) \iff \boldsymbol{\mathcal{M}}_\alpha(w) \models \varphi$$

$$\mathbf{p}_{\alpha+1}(w)(\varphi) = m_w\{v \mid \boldsymbol{\mathcal{M}}_\alpha(v) \models \varphi\}$$

and limit stages such that:

If $\mathrm{T}(\varphi) = 1$ is nearly stable in $\langle\boldsymbol{\mathcal{M}}_\alpha(w)\rangle_{\alpha<\mu}$ then $\varphi \in \mathbf{T}_\mu(w)$

If $r \leqslant \mathrm{p}(\varphi) \leqslant q$ is nearly stable in $\langle\boldsymbol{\mathcal{M}}_\alpha(w)\rangle_{\alpha<\mu}$ then $r \leqslant \mathbf{p}_\mu(w)(\varphi) \leqslant q$

And $\mathbf{T}_\mu(w)$ is a maximally consistent set of sentences and $\mathbf{p}_\mu(w)$ a finitely additive probability[30] then:

The axiomatic theory $\mathsf{PT}_2'^{\mathsf{PMS}}$ is satisfied at all limit ordinals, where $\mathsf{PT}_2'^{\mathsf{PMS}}$ modifies PT_2' by replacing $\mathsf{Prov}_{\mathsf{PT}_1}$ with $\mathsf{Prov}_{\mathsf{PT}_1^{\mathsf{PMS}}}$ and also dropping Probabilistic Convention T (unless the limit was designed to satisfy that, for example as given by BanLim) and the version of Miller's Principle.

If near stability is replaced with stability, $\mathsf{PT}_2'^{\mathsf{PMS}}$ will be satisfied at limit ordinals $\alpha + \omega$.

We can also add extra axioms in the probabilistic modal structure case depending on the structure on which the construction is based. For example principles for introspection or trust. We discuss these in Section 5.4.2.

---

[29] "$\forall\varphi$" is understood as $\forall x(\mathsf{Sent}_{\mathsf{P},\mathsf{T}}(x) \to \ldots)$. Quantification into $\ulcorner \urcorner$ can be made precise using standard arithmetic means. For further details of a way to do this see Halbach (2014).

[30] Such as the notions of revision sequences discussed in Sections 5.3.2 and 5.3.3.

These theories we have given are sound but carry no completeness property. However, they provide a start to reasoning about the constructions syntactically. Since we have been working on giving nice limit stages we see that the theory of the limit stages, as is given in PT_2' and $\mathsf{PT}_2'^{\mathsf{PMS}}$, are in fact interesting theories. This is different to the usual revision sequence where the limit stages are not intended as interpretations for the truth predicate but are instead tools, for example they don't have $\varphi \notin \mathrm{T}_\mu \iff \neg\varphi \in \mathrm{T}_\mu$. This alternative focus might lead to very interesting models arising from these constructions which was ignored in the usual constructions where the limit condition is designed to be fairly weak.

5.4.2   Further conditions we could impose

The importance of the construction based on the probabilistic modal structures is that it allows us to consider different probabilistic modal structures which will have implications for the interactions between the probabilities. By considering particular probabilistic modal structures we can add extra axioms.

Introspection

An important class of probabilistic modal structures are the introspective ones. Strongly introspective frames are the ones where:[31]

$$m_w\{v \mid m_v = m_w\} = 1$$

In this construction we should modify the expression of introspection in an analogous way to as we did in Section 3.4.1.

Proposition 5.4.5. *If $\mathfrak{M}$ is weakly introspective,[32] and $\boldsymbol{\mathcal{M}}_{\alpha+1}$ is given by*

$$\varphi \in \mathbf{T}_{\alpha+1}(w) \iff \boldsymbol{\mathcal{M}}_\alpha \models \varphi$$

$$\mathbf{p}_{\alpha+1}(w)(\varphi) = m_w\{v \mid \boldsymbol{\mathcal{M}}_\alpha(v) \models \varphi\},$$

we will have:[33]

$$\boldsymbol{\mathcal{M}}_\alpha(w) \models \mathsf{T}\ulcorner \mathsf{P}\ulcorner\varphi\urcorner \geqslant \bar{r}\urcorner \to \mathsf{P}\ulcorner \mathsf{P}\ulcorner\varphi\urcorner \geqslant \bar{r}\urcorner = 1$$
$$\boldsymbol{\mathcal{M}}_\alpha(w) \models \mathsf{T}\ulcorner \neg\mathsf{P}\ulcorner\varphi\urcorner \geqslant \bar{r}\urcorner \to \mathsf{P}\ulcorner \neg\mathsf{P}\ulcorner\varphi\urcorner \geqslant \bar{r}\urcorner = 1$$

for all $\alpha > 1$, and therefore this has probability 1 in the limit probability.

---

[31] In the σ-additive case these are called Harsanyi type spaces. For applications of these spaces, this assumption is often taken for granted.

[32] At least for agent A. This is a slight weakening of the condition $m_w\{v \mid m_v = m_w\} = 1$.

[33] Also note that the negative form is much stronger than the version in the Kripkean construction, Section 3.4.1, because in this theory we have $\mathsf{T}\ulcorner\neg\varphi\urcorner \leftrightarrow \neg\mathsf{T}\ulcorner\varphi\urcorner$ and $\mathsf{P}\ulcorner\neg\varphi\urcorner = 1 - \mathsf{P}\ulcorner\varphi\urcorner$. Also because of the ability to switch the order of T and P one could equivalently write these with $\mathsf{P}\ulcorner\mathsf{T}\ulcorner\varphi\urcorner\urcorner \geqslant r$ or $\neg\mathsf{P}\ulcorner\mathsf{T}\ulcorner\varphi\urcorner\urcorner \geqslant r$ as antecedents.

Proof. We will only present the proof for the positive version.

$$\boldsymbol{\mathcal{M}}_{\alpha+2}(w) \models \mathsf{T}^{\ulcorner}\mathsf{P}^{\ulcorner}\varphi^{\urcorner} \geqslant \bar{r}^{\urcorner}$$
$$\iff \boldsymbol{\mathcal{M}}_{\alpha+1}(w) \models \mathsf{P}^{\ulcorner}\varphi^{\urcorner} \geqslant \bar{r}$$
$$\iff m_w\{u \mid \boldsymbol{\mathcal{M}}_{\alpha}(u) \models \varphi\} \geqslant r$$
$$\implies m_w\{v \mid m_v\{u \mid \boldsymbol{\mathcal{M}}_{\alpha}(u) \models \varphi\} \geqslant r\} = 1 \quad \mathfrak{M} \text{ weakly introspective}$$
$$\implies m_w\{v \mid \boldsymbol{\mathcal{M}}_{\alpha+1}(v) \models \mathsf{P}^{\ulcorner}\varphi^{\urcorner} \geqslant r\} = 1$$
$$\iff \boldsymbol{\mathcal{M}}_{\alpha+2}(w) \models \mathsf{P}^{\ulcorner}\mathsf{P}^{\ulcorner}\varphi^{\urcorner} \geqslant r^{\urcorner} = 1$$

$\square$

The addition of the truth predicate in the expression of introspection prevents the contradiction from arising. We can give an explanation for this as described in Fig. 5.5.
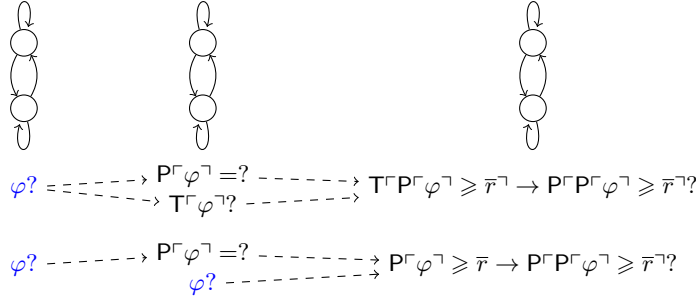
Figure 5.5: Why introspection formulated this way is consistent

What this is pointing out is that the answer to whether the introspection version using T is satisfied at $\boldsymbol{\mathcal{M}}_{\alpha+2}(w)$ depends only on the question of at which $u \in W$, $\boldsymbol{\mathcal{M}}_{\alpha}(u) \models \varphi$ and then the successor definition and the underlying $\mathfrak{M}$ do the rest of the work. However, the version not using T depends both on the question of at which $u \in W$, $\boldsymbol{\mathcal{M}}_{\alpha}(u) \models \varphi$ and where $\boldsymbol{\mathcal{M}}_{\alpha+1}(u) \models \varphi$. How these answers relate depends on what φ is. But for example if $\varphi = \mathsf{T}^{\ulcorner}\psi^{\urcorner}$, then the answer to whether $\boldsymbol{\mathcal{M}}_{\alpha}(u) \models \varphi$ will be the opposite answer to whether $\boldsymbol{\mathcal{M}}_{\alpha+1}(u) \models \varphi$. So these answers need not cohere, and the point where they are put together in the conditional then might have strange features. In particular then it does not just depend on the structure of $\mathfrak{M}$ (and, of course, the successor definition).

We shall now move to considering the requirement of deference and shall see the same thing happening.

Reformulating deference

We showed in Section 1.7 that a deference principle

$$\mathsf{P}^A(^{\ulcorner}\varphi^{\urcorner} \mid {}^{\ulcorner}\mathsf{P}^B{}^{\ulcorner}\varphi^{\urcorner} \geqslant 1/2^{\urcorner}) \geqslant 1/2$$
$$\mathsf{P}^A(^{\ulcorner}\varphi^{\urcorner} \mid {}^{\ulcorner}\neg\mathsf{P}^B{}^{\ulcorner}\varphi^{\urcorner} \geqslant 1/2^{\urcorner}) \ngeqslant 1/2$$

is problematic in a framework where there are self-referential probabilities.

But using our considerations as for introspection we can find an alternative version of these principles which is consistent by using the truth predicate.

Proposition 5.4.6. *The following are consistent:*

- P *is probabilistic over a base theory of arithmetic, or more generally* $\mathsf{PT}_1^{\mathsf{PMS}}$ *holds*

- *Trust formulated with a* T *predicate:*

  - $\forall\varphi\forall a \in R(\mathsf{P}^{\mathrm{A}}(\ulcorner\mathsf{P}^{\mathrm{B}}\ulcorner\varphi\urcorner \geqslant a\urcorner) > 0 \rightarrow \mathsf{P}^{\mathrm{A}}(\ulcorner\mathsf{T}\ulcorner\varphi\urcorner\urcorner|\ulcorner\mathsf{P}^{\mathrm{B}}\ulcorner\varphi\urcorner \geqslant a\urcorner) \geqslant a)$
  - $\forall\varphi\forall a \in R(\mathsf{P}^{\mathrm{A}}(\ulcorner\neg\mathsf{P}^{\mathrm{B}}\ulcorner\varphi\urcorner \geqslant a\urcorner) > 0 \rightarrow \neg\mathsf{P}^{\mathrm{A}}(\ulcorner\mathsf{T}\ulcorner\varphi\urcorner\urcorner \mid \ulcorner\neg\mathsf{P}^{\mathrm{B}}\ulcorner\varphi\urcorner \geqslant a\urcorner) \geqslant a)$
  - *And similarly for* $=, >, <, \leqslant$ *and also intervals like* $a < \mathsf{P}^{\mathrm{B}}\ulcorner\varphi\urcorner \leqslant b$.

Proof. One can show that in a frame where $m_w^{\mathrm{A}}\{v \mid m_v^{\mathrm{B}} = m_w^{\mathrm{A}}\} = 1$, we have for $n > 1$

$$\boldsymbol{\mathcal{M}}_\alpha(w) \models \mathsf{P}^{\mathrm{A}}\ulcorner\mathsf{T}\ulcorner\varphi\urcorner\urcorner = \bar{r} \rightarrow \mathsf{P}^{\mathrm{A}}\ulcorner\mathsf{P}^{\mathrm{B}}\ulcorner\varphi\urcorner = \bar{r}\urcorner = 1$$

and the deference principles we require are consequences of this. $\qquad\square$

In fact, deference formulated with a truth predicate, in the setting where we formulate P as a function symbol is satisfied in exactly the frames where the operator variants without T are satisfied. Though I am currently unaware of the frame condition for deference, this would be a result analogous to Proposition 2.3.2.

As for the case of introspection, the explination for why this does not lead to contradiction can be summed up in Fig. 5.6:
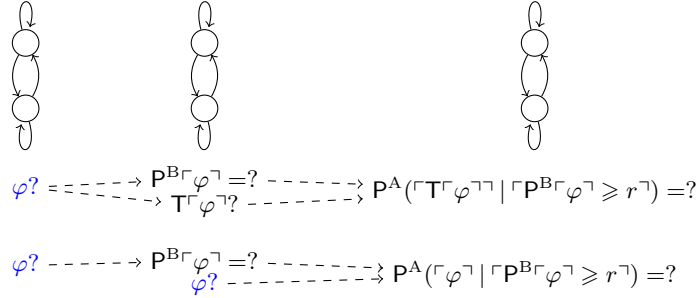
Figure 5.6: Why deference formulated this way is consistent

5.5 Conclusion

We have presented a number of possibilities for how to define a revision sequence to also account for probability. We were interested in giving a limit criterion which leads to nice limit stages, so we presented a criterion which gives as much as possible in the limit stages in the sense that we consider as many properties of the extensions of T and P as possible. We gave two different ways that the successor probabilities can be determined, the first was to take

relative frequencies over the previous stages, the second is to use additional structure in the background. Both of these constructions are interesting in their own right. The former style of revision sequence has that the feature that the probability notion now tells us useful extra information about truth and other logical facts. For example in those sequence one will always have that $p_\mu(\lambda) = 1/2$. The second style of revision sequence, however, can apply to our everyday use of probabilities as one can allow for varying interpretations of the probability notion. Though we have not come to definite conclusions in this chapter we have demonstrated that there are lots of possibilities for revision theories of truth *and probability* and we hope that these may lead to valuable insights into the concept of probability as it applies in expressively rich settings.

In this chapter we have presented a new limit stage definition that is different from traditional revision sequences. We have been able to propose such a limit stage because the notion that we are trying to determine is probability, which takes values in the real numbers. In the revision theory we are interested in what is brought about by the stages beneath some limit stage. Since for truth there are only two truth values, the only features that can be *brought about* is to end up being equal to one of these truth values. Now that we instead have the real numbers to provide values, we might instead have that the probability of a sentence *converges* to some value and then use this information to let the limit probability be the limit of these values. This motivated us to consider generalising the notion of the limit stage. We were then naturally lead to further generalisations by considering even more properties, for example also those stating relationships between different sentences. The same advantages may be had with something like fuzzy truth which would also have a domain of $\mathbb{R}$, but the application to probability is interesting because of the usefulness of probability. We can also take insights from this construction and focus just on how truth is revised. In our revision sequences, equivalences of liars were respected at limits, which is an interesting property that hasn't previously been studied.

Appendix 5.A   Definition of closed

Here we will prove the result from Proposition 5.2.5. However, we will state it in the more general way where the language may be uncountable, typically because it would then have constants for all real numbers. Then one would need to work with an alternative syntax coding, but the details of this won't matter for our purposes.

Proposition 5.A.1. *Suppose $C \subseteq \mathsf{Mod}$ depends on only countably many sentences, i.e. There is some countable collection of sentences $\{\varphi_n\}_{n \in \mathbb{N}}$ such that if for all n, $p(\varphi_n) = p'(\varphi_n)$ and $\varphi_n \in \mathrm{T} \leftrightarrow \varphi_n \in \mathrm{T}'$, then $(M, p, \mathrm{T}) \in C \iff (M, p', \mathrm{T}') \in C$.*

Then the following are equivalent:

1. *$C \subseteq \mathsf{Mod}$ is closed*

2. *For every sequence (M, p_n, T_n) and each $(\mathrm{M}, p_{lim}, \mathrm{T}_{lim})$ such that $(\mathrm{M}, p_n, \mathrm{T}_n) \longrightarrow (\mathrm{M}, p_{lim}, \mathrm{T}_{lim})$, i.e.*

   - *for every φ $p_n(\varphi) \longrightarrow p_{lim}(\varphi)$[34]*

---

[34]I.e. for all $\epsilon \in \mathbb{R}_{>0}$, there is some $m \in \mathbb{N}$ such that for all $n > m$, $|p_n(\varphi) - p_{lim}(\varphi)| < \epsilon$.

- *for every φ there is some $m \in \mathbb{N}$ such that either*
  - *for every $n > m$, $\varphi \in \mathrm{T}_n$ and $\varphi \in \mathrm{T}_{lim}$, or*
  - *for every $n > m$, $\varphi \notin \mathrm{T}_n$ and $\varphi \notin \mathrm{T}_{lim}$.*

  If each $(\mathrm{M}, \mathrm{p}_n, \mathrm{T}_n) \in C$, then $(\mathrm{M}, \mathrm{p}_{lim}, \mathrm{T}_{lim}) \in C$.

Proof. For *Item 1 $\implies$ Item 2* work by induction on the definition of being closed, observing that all members of the subbase have this property, and the inductive definition steps respect the property.

For *Item 2 $\implies$ Item 1*: Suppose C only depends on $\{\varphi_n\}$ and Item 2 is satisfied. Define

$$\mathsf{Mod}\!\restriction_{\{\varphi_n\}} := \bigcap \{\mathcal{M} \in \mathsf{Mod} \mid \forall \psi \notin \{\varphi_n\},\ \psi \notin \mathrm{T} \text{ and } \mathrm{p}(\psi) = 0\}$$

and endow it with the subspace topology. Observe that this is topologically equivalent to $\mathsf{Mod}_{\mathcal{L}} \times \mathbb{R}^{\{\varphi_n\}} \times \{0,1\}^{\{\varphi_n\}}$ with the product topology (where $\mathsf{Mod}_{\mathcal{L}}$ has the discrete topology). So it is first countable and therefore sequential. By the assumption that Item 2 is satisfied we have that C is sequentially closed in $\mathsf{Mod}\!\restriction_{\{\varphi_n\}}$ and therefore that it is closed in the topology on $\mathsf{Mod}\!\restriction_{\{\varphi_n\}}$. Since $\mathsf{Mod}\!\restriction_{\{\varphi_n\}}$ is a closed subset of Mod we have that C is closed in Mod too. $\qquad\square$

Appendix 5.B Proof that there are infinitely many choices at limit stages

In this section we include the proof of Theorem 5.2.18.

Lemma 5.B.1. *Let $\langle (\mathrm{M}, \mathrm{T}_\alpha, \mathrm{p}_\alpha) \rangle_{\alpha < \mu} = \langle \mathcal{M}_\alpha \rangle_{\alpha < \mu}$ be any μ-length sequence of members of Mod. Suppose A is a closed set that is nearly cofinal beneath μ in $\langle \mathcal{M}_\alpha \rangle_{\alpha < \mu}$ i.e. such that $\mathsf{Mod} \setminus A$ is not nearly stable beneath μ. Then we can find some $(\mathrm{M}, \mathrm{T}_\mu, \mathrm{p}_\mu) = \mathcal{M}_\mu \in A$ such that:*

> *Whenever C is a closed property that is nearly stable in $\langle \mathcal{M}_\alpha \rangle_{\alpha < \mu}$ then $\mathcal{M}_\mu \in C$.*

This is just a slight generalisation of Theorem 5.2.11. It can be proved by choosing some $\alpha \geqslant \max\{\beta^i \mid i \in \{1, \dots, m\}\}$ and $k \geqslant \max\{N_\alpha^i \mid i \in \{1, \dots, m\}\}$ where $\mathcal{M}_{\alpha+k} \in A$ since therefore this is in $A \cap C_1 \cap \dots \cap C_n$, and so $A \cap \bigcap \mathcal{C} \neq \varnothing$.

Proof. Let

$$\mathcal{C} := \{C \text{ closed} \mid C \text{ is nearly stable beneath } \mu \text{ in } \langle \mathcal{M}_\alpha \rangle_{\alpha < \mu}\}$$

We will show that $\mathcal{C} \cup \{A\}$ has the finite intersection property. We will then be able to deduce that

$$\bigcap (\mathcal{C} \cup \{A\}) \neq \varnothing$$

which suffices to prove the theorem since taking some $\mathcal{M}_\mu \in \bigcap(\mathcal{C} \cup \{A\})$ will be as required. Note that this suffices because $\{(\mathrm{M}', \mathrm{T}, \mathrm{p}) \in \mathsf{Mod} \mid \mathrm{M}' = \mathrm{M}\}$ is closed.

5.B Proof that there are infinitely many choices at limit stages

Let $\mathcal{D}$ be finite $\subseteq \mathcal{C} \cup \{A\}$. Enumerate the members of $\mathcal{D}$ so that $\mathcal{D} \cup \{A\} = \{C_1, \ldots, C_m, A\}$.[35]

We know that for each $i = 1, \ldots, n$

$$\exists \beta^i < \mu \; \forall \alpha^{<\mu}_{>\beta^i} \; \exists N^i_\alpha < \omega \; \forall n^{<\omega}_{>N^i_\alpha} \quad \mathcal{M}_{\alpha+n} \in C_i$$

Therefore for each $\alpha \geqslant \max\{\beta^i \mid i \in \{1, \ldots, m\}\}$ and $k \geqslant \max\{N^i_\alpha \mid i \in \{1, \ldots, m\}\}$

$$\mathcal{M}_{\alpha+k} \in C_1 \cap \ldots \cap C_m$$

$\mathsf{Mod} \setminus A$ is not nearly stable beneath μ in $\langle \mathcal{M}_\alpha \rangle_{\alpha < \mu}$, so:

$$\forall \beta < \mu \; \exists \alpha^{<\mu}_{>\beta} \; \forall N_\alpha < \omega \; \exists n^{<\omega}_{>N_\alpha} \quad \mathcal{M}_{\alpha+n} \in A$$

Therefore there is some α with $\mu > \alpha > \max\{\beta^i \mid i \in \{1, \ldots n\}\}$ and n with $\omega > n > \max\{N^i_\alpha \mid i \in \{1, \ldots, m\}\}$ which is such that $\mathcal{M}_{\alpha+n} \in A$. By the previous observation this will also be a member of $C_1 \cap \ldots \cap C_m$.

So we can see that $\bigcap(\mathcal{D} \cup \{A\}) = \bigcap\{C_1, \ldots, C_m, A\} \neq \varnothing$, so $\bigcap \mathcal{D} \neq \varnothing$. We have therefore shown that $\mathcal{C} \cup \{A\}$ has the finite intersection property, as required. $\qquad \square$

Lemma 5.B.2. *Let $\langle \mathcal{M}_\alpha \rangle_{\alpha < \mu}$ be any μ-length initial sequence of a revision sequence in the sense of Definition 5.2.8. Consider $\delta \leftrightarrow (\mathsf{P}^\ulcorner \delta^\urcorner \leqslant 0.4 \vee (0.4 < \mathsf{P}^\ulcorner \delta^\urcorner < 0.6 \wedge \mathsf{T}^\ulcorner \delta^\urcorner))$, as in* IN AN INTERVAL *from Example 5.2.9. We see that for every M and $m < M$*

$$0.4 + 0.2 \cdot \frac{m}{M} \leqslant \mathsf{P}^\ulcorner \delta^\urcorner \leqslant 0.4 + 0.2 \cdot \frac{m+1}{M}$$

has its complement not nearly stable beneath μ.

Proof. We prove this by a series of lemmas.

Lemma 5.B.3. *For every φ, μ and n:*

$$\mathrm{p}_{\mu+n+1}(\varphi) = \mathrm{p}_{\mu+n}(\varphi) + \frac{[\![\varphi]\!]_{\mathcal{M}_{\mu+n}} - \mathrm{p}_{\mu+n}(\varphi)}{n+1}$$

So, if $[\![\varphi]\!]_{\mathcal{M}_{\mu+n}} = 1$ then $\mathrm{p}_{\mu+n+1}(\varphi) \geqslant \mathrm{p}_{\mu+n}(\varphi)$, otherwise $\mathrm{p}_{\mu+n+1}(\varphi) \leqslant \mathrm{p}_{\mu+n}(\varphi)$

Proof. Simple manipulation:

$$\mathrm{p}_{\mu+n+1}(\varphi) = \frac{[\![\varphi]\!]_{\mathcal{M}_\mu} + \ldots + [\![\varphi]\!]_{\mathcal{M}_{\mu+n-1}} + [\![\varphi]\!]_{\mathcal{M}_{\mu+n}}}{n+1}$$

$$= \frac{n \cdot \frac{[\![\varphi]\!]_{\mathcal{M}_\mu} + \ldots + [\![\varphi]\!]_{\mathcal{M}_{\mu+n-1}}}{n} + [\![\varphi]\!]_{\mathcal{M}_{\mu+n}}}{n+1}$$

$$= \frac{n \cdot \mathrm{p}_{\mu+n}(\varphi) + [\![\varphi]\!]_{\mathcal{M}_{\mu+n}}}{n+1}$$

$$= \frac{(n+1)\mathrm{p}_{\mu+n}(\varphi) - \mathrm{p}_{\mu+n}(\varphi) + [\![\varphi]\!]_{\mathcal{M}_{\mu+n}}}{n+1}$$

$$= \mathrm{p}_{\mu+n}(\varphi) + \frac{[\![\varphi]\!]_{\mathcal{M}_{\mu+n}} - \mathrm{p}_{\mu+n}(\varphi)}{n+1}$$

---

[35]We shall show that $\bigcap(\mathcal{D} \cup \{A\}) \neq \varnothing$ so we do not have to deal with the separate cases where $A \in \mathcal{D}$ and $A \notin \mathcal{D}$.

$\square$

Lemma 5.B.4. *For each* M *and* $N < \omega$ *we can pick* $\omega > k > N$ *where for every* $n > k$, $|p_{\alpha+n}(\delta) - p_{\alpha+n+1}(\delta)| \leqslant \frac{0.2}{M}$

Proof. $|p_{\alpha+n}(\delta) - p_{\alpha+n+1}(\delta)| \leqslant \max\{|\frac{1-p_{\mu+n}(\delta)}{n+1}|, |-\frac{p_{\mu+n}(\delta)}{n+1}|\} \leqslant \frac{1}{n+1}$.
By picking $k \geqslant \max\{5M, N\}$, we have $|p_{\alpha+n}(\delta) - p_{\alpha+n+1}(\delta)| \leqslant \frac{1}{5M+1} \leqslant \frac{0.2}{M}$, as required. $\square$

Lemma 5.B.5. *For each* $\alpha < \mu$ *and* $N < \omega$ *we can pick* $\omega > n > N$ *where* $p_{\alpha+n} \leqslant 0.4$ *and* $\omega > n' > N$ *where* $p_{\alpha+n} \geqslant 0.6$.

Proof. Suppose $[\![\delta]\!]_{\mathcal{M}_{\alpha+N}} = 1$. Then $p_{\alpha+N}(\delta) < 0.6$ by the definition of δ. So the probability of δ will keep increasing until we reach an n where $p_{\alpha+n}(\delta) \geqslant 0.6$. For each k with $p_{\alpha+N+k}(\delta) < 0.6$, $p_{\alpha+N+k}(\delta) = \frac{Np_{\alpha+N}(\delta)+k}{N+k}$. This can be proved by induction on k: Base case $k = 0$ is clear. For $k+1$, if $p_{\alpha+N+k+1}(\delta) < 0.6$ then:

$$p_{\alpha+N+k+1+1}(\delta) = p_{\alpha+N+k+1}(\delta) + \frac{1 - p_{\alpha+N+k+1}(\delta)}{N+k+1}$$

$$= \frac{Np_{\alpha+N}(\delta)+k}{N+k} + \frac{1 - \frac{Np_{\alpha+N}(\delta)+k}{N+k}}{N+k+1}$$

$$= \frac{(N+k+1)(Np_{\alpha+N}(\delta)+k) + (N+k) - (Np_{\alpha+N}(\delta)+k)}{(N+k)(N+k+1)}$$

$$= \frac{(N+k)(Np_{\alpha+N}(\delta)+k) + (N+k)}{(N+k)(N+k+1)}$$

$$= \frac{Np_{\alpha+N}(\delta)+k+1}{N+k+1}$$

Now $\frac{Np_{\alpha+N}(\delta)+2N}{N+2N} \geqslant \frac{2N}{3N} \geqslant 0.6$, so there will be some $k \leqslant 2N$ with $p_{\alpha+N+k}(\delta) \geqslant 0.6$.

Now suppose $[\![\delta]\!]_{\mathcal{M}_{\alpha+N}} = 0$. We apply analogous reasoning to show that while $p_{\alpha+N+k}(\delta) > 0.4$, $p_{\alpha+N+k}(\delta) = \frac{Np_{\alpha+N}(\delta)}{N+k}$, and see that $\frac{Np_{\alpha+N}(\delta)}{N+2N} \leqslant \frac{N}{3N} \leqslant 0.4$.

If $[\![\delta]\!]_{\mathcal{M}_{\alpha+N}} = 1$ we have found our $k > N$ where $p_{\alpha+N+k} \leqslant 0.4$ so we can take $n = N + k$, then use the second argument with $N' = N + k$ to find $0 < k' \leqslant 2N'$ where $p_{\alpha+N'+k'} \geqslant 0.6$, so can take $n = N'+k'$. If $[\![\delta]\!]_{\mathcal{M}_{\alpha+N}} = 0$ we do the process the other way around, first finding one with high enough probability then one with low enough. $\square$

Corollary 5.B.6. *For any* $\alpha < \mu$ *and* N, *and* M *we can pick* $n > N$ *so that*

$$0.4 + 0.2 \cdot \frac{m}{M} \leqslant p_{\alpha+n}(\delta) \leqslant 0.4 + 0.2 \cdot \frac{m+1}{M}$$

Proof. Choose k where for each $n > k$, $|p_{\alpha+n}(\delta) - p_{\alpha+n+1}(\delta)| \leqslant \frac{0.2}{M}$ using Lemma 5.B.4 Choose $k' > k$ with $p_{\alpha+k}(\delta) \leqslant 0.4$ using Lemma 5.B.5[36] and $k'' > k$ with $p_{\alpha+k''}(\delta) \geqslant 0.6$ again using Lemma 5.B.5.

---

[36]The previous lemmas were stated for μ and this one for arbitrary α, so if α is a successor ordinal apply Lemma 5.B.5 to find $n > N + k_\alpha$ such that $p_{\zeta_\alpha+n}(\delta) \leqslant 0.4$ and take $k' = n - k_\alpha$.

$p_{\alpha+k'}(\delta) \leqslant 0.4$ and $p_{\alpha+k''}(\delta) \geqslant 0.6$ and the probability values move from 0.4 to 0.6 in small enough jumps (because $k' > k$), the probability values must lie within each $1/M$-sized interval between 0.4 and 0.6. More carefully we can argue as follows:

Prove by induction on n_0 that for all $n_0 \geqslant k'$, if there is no n with $k' \leqslant n \leqslant n_0$ such that $0.4 + 0.2 \cdot \frac{m}{M} \leqslant p_{\alpha+n}(\delta) \leqslant 0.4 + 0.2 \cdot \frac{m+1}{M}$ then for every n with $k' \leqslant n \leqslant n_0$, $p_{\alpha+n}(\delta) \leqslant 0.4 + 0.2 \cdot \frac{m}{M}$.

For $n_0 = k'$ the result is clear.

Suppose there is no n such that $k' \leqslant n \leqslant n_0 + 1$ and $0.4 + 0.2 \cdot \frac{m}{M} \leqslant p_{\alpha+n}(\delta) \leqslant 0.4 + 0.2 \cdot \frac{m+1}{M}$. Then by the induction hypothesis for every n with $p_{\alpha+n_0}(\delta) \leqslant 0.4 + 0.2 \cdot \frac{m}{M}$, so since $|p_{\alpha+n_0}(\delta) - p_{\alpha+n_0+1}(\delta)| \leqslant \frac{0.2}{M}$, $p_{\alpha+n_0+1}(\delta) \leqslant p_{\alpha+n_0}(\delta) + \frac{0.2}{M} \leqslant 0.4 + 0.2 \cdot \frac{m+1}{M}$. Since we assumed that $n_0 + 1$ does not have the property that $0.4 + 0.2 \cdot \frac{m}{M} \leqslant p_{\alpha+n}(\delta) \leqslant 0.4 + 0.2 \cdot \frac{m+1}{M}$, it must be that $p_{\alpha+n_0}(\delta) < 0.4 + 0.2 \cdot \frac{m}{M}$.

Now since we know that there is a k'' where $p_{\alpha+k'} \geqslant 0.6$, we have for each $m < M$ this k' is such that $p_{\alpha+k'} \geqslant 0.4 + 0.2 \cdot \frac{m}{M}$. And therefore there is an n with $k' \leqslant n \leqslant k''$ such that $0.4 + 0.2 \cdot \frac{m}{M} \leqslant p_{\alpha+n}(\delta) \leqslant 0.4 + 0.2 \cdot \frac{m+1}{M}$ by using the contrapositive of what we just proved by induction. This is as required. $\qquad\square$

$\square$

Part II

Rationality Requirements

Chapter 6

Introduction

In this second part of the thesis we will turn to a different, but related, question:

> What rationality requirements are there on agents in such expressively rich frameworks?

and, relatedly:

> To what degree should an agent believe a sentence that says something about her own degrees of belief?

In Part I we developed semantics for frameworks that involve such self-referential sentences, but in this part we consider how traditional arguments for rational requirements on agents, such as probabilism, apply when such self-referential sentences are around. In this we are therefore focusing on the particular interpretation of probability as subjective probability, or degrees of belief of an agent.

6.1   The question to answer

There has been a large body of work trying to develop justifications for particular rationality constraints on agents, particularly focused on justifying probabilism. There are two main influential styles of argument: an argument from *accuracy*, initially presented in Joyce (1998), and a so-called *Dutch book argument*, originating from Ramsey (1931). The argument from accuracy says an agent should have credences that are as *accurate* as possible. The Dutch book argument says that agents should have credences which, if they bet in accordance with these credences, will not lead them to a guaranteed loss of money. It is not yet clear that the semantics that we have proposed can model agents who are doing best from an accuracy or Dutch book point of view. This is the question that we turn to in this part of the thesis.

Michael Caie has recently argued (Caie, 2013) that accuracy and Dutch book criteria need to be modified if there are self-referential probabilities, and that appropriately modified they in fact lead to the requirement that a rational agent must have degrees of belief which are *not* probabilistic and which are also not representable in any of the semantics we have proposed in Part I. If it turned out that Caie's suggested modifications of the criteria were correct, then this

would be a blow to our proposed semantics. Perhaps the appropriate response in that case would be to admit that our semantics are unable to model rational agents, so the notion of probability embedded in these semantics could not be interpreted as subjective probability.

Caie's suggested modification of the accuracy criterion is that we should consider the inaccuracy of the *act of coming to occupy* a credal state *c*, analysing this in a decision-theoretic manner. To do this we only care about the inaccuracy of a credal state at the world which would be actual if the agent had those credences. In cases where we have sentences where having some attitude towards them can affect their truth value, it turns out that this then differs from the traditional accuracy criterion.

In Chapter 7 we will consider Caie's suggested modification and we will show a number of undesirable consequences of it. These will give us more motivation to consider rejecting his modification and instead consider something analogous to the usual accuracy criterion. This leaves open the possibility that accuracy considerations do in fact support the semantics we provided in Part I, and in Section 7.3.2 we will show that one way of understanding the accuracy criterion does in fact lead to the semantics developed. In doing this we will still need some additional generalisations and considerations in formulating the accuracy criterion because the semantics we developed, at least in Chapters 3 and 4, dropped certain assumptions implicit in the traditional accuracy criterion by dropping classical logic and the assumption that credences assign single real numbers to each sentence. We will briefly consider how one might apply these considerations in such a setting in Section 7.3.4. This connects to work by J. Robert G. Williams (2012b; 2014) on non-classical probabilities. In the semantics developed in those chapters we were able to find *fixed points*, which in Chapter 4 we called *stable states*. It will turn out that for the way we suggest to formulate the rational constraints in Section 7.3.4, these will be exactly the credences that are *immodest*, or look the best from their own perspective, so these are desirable credal states.

In Chapter 8 we will consider the Dutch book argument and will work with the assumption that an agent in such a situation does bet in accordance with her credences, and under that assumption try to develop a Dutch book criterion which is applicable in a wide range of circumstances. In developing this criterion we are expanding a suggestion from Caie (2013). We will show that the proposal that we finally settle on is in fact a version of the modified accuracy criterion that we considered in Chapter 7. It therefore inherits a number of undesirable characteristics. This will therefore lend more weight to our proposal to in fact reject this criterion by rejecting the assumption that an agent bet with her credences.

One problem before we can even get started with considering these arguments is what the worlds at stake are. Both the accuracy and Dutch book arguments, at least as traditionally formulated, refer to the notion of a collection of possible worlds. The accuracy criterion can be formulated as: A credal state is irrational if there is some alternative credal state which is more accurate *whatever the world is like*. The Dutch book criterion can be formulated as: A credal state is irrational if there is a bet that the agent would be willing to accept but will lead her to a loss *whatever the world is like*. Caie already suggests that we should not consider *any* way the world would be like, but only those that are consistent with her having the considered credences. But there is also a

more fundamental worry: what *could* the world be like? Surely that's governed by a *semantics*, so we would need to determine a semantics before we can even discuss these arguments. However, in the very simple case that Caie considers, such an analysis of a potential semantics is not a prerequisite for discussing the arguments. In Caie's analysis he assumes that agents assign point-valued degrees of belief to each sentence and that the background semantics is classical. In the cases that he considers, the agent's beliefs in empirical matters are irrelevant, so we can also consider the agents as omniscient. Furthermore he considers the agent to be introspective. In the possible worlds framework we can see that he is considering the trivial probabilistic modal structure $\mathfrak{M}_{\mathsf{omn}}$. So he considers situations where *if the agent's (A's) credences are c, then c is the correct interpretation of* P^{A} *in the object language, and the agent is herself aware of this.* So for his considerations we do not need a more developed semantics.

However, when the assumptions of introspection and omniscience are dropped then we need to do something else. For example we will then consider the agents as modelled by other probabilistic modal structures and determine what their degrees of belief should be like. This will then be closely related to the revision semantics considered in Chapter 5 and will be discussed in Section 7.3.3. In Section 7.3.4 we will also consider dropping the assumption that an agent have point-valued probabilities and that the background logic is classical.

6.2 Setup

In this part of the thesis we will work with $\mathcal{L}_{\mathsf{P}}$, for $\mathcal{L}$ any language extending $L_{\mathsf{PA,ROCF}}$, which formalises the probability notion as a function symbol. However, these details of the language will not be important for our discussion. For example in Theorem 7.1.11 we will consider a more expressive language.

We will be judging an agent's credal state by looking at a specific "agendas", or collections of sentences that we are interested in.

Definition 6.2.1. An *agenda,* $\mathcal{A}$, is some collection of sentences. I.e. $\mathcal{A} \subseteq \mathsf{Sent}_{\mathsf{P}}$.

We will typically fix an agenda and judge the agent's rationality just by considering her credences in the sentences in that agenda. This is ideally used as a tool to give us some sentences to focus on.[1] Unless otherwise stated, an agenda will be assumed to be finite.

We will use φ as a metavariable for arbitrary sentences, and δ for a metavariable for sentences of the form:

$$\delta \leftrightarrow \text{`the agent's credences, } c, \text{ are} \in R\text{'}$$

Note, here the quote marks are scare-quotes. One would replace that part of the sentence with some sentence in the formal language describing the fact that

---

[1]We would want a result to show that this focusing on agendas is not important in the following sense:

Desired Theorem 6.2.2. *If an agent is rational judged with respect to agenda* $\mathcal{A}$, *then she is rational when judged with respect to agenda* $\mathcal{A}' \subseteq \mathcal{A}$.

In fact all the rationality criteria which we consider do satisfy this theorem.

her credences satisfy the relevant constraint. E.g. $\mathsf{P}\ulcorner\varphi_0\urcorner \geqslant {}^1\!/_2$. Throughout this section we will often just state a biconditional to characterise the sentence. We will be assuming that this biconditional is derivable in PA.[2] δ is the sort of sentence we would obtain by the diagonal lemma, so it might be e.g.

$$\delta \leftrightarrow \neg\mathsf{P}\ulcorner\delta\urcorner \geqslant {}^1\!/_2$$

though this special case of δ is called π. Or,

$$\delta \leftrightarrow (\mathsf{P}\ulcorner\delta\urcorner \leqslant 0.5 \vee (\mathsf{P}\ulcorner\delta\urcorner \leqslant 0.55 \wedge \mathsf{P}\ulcorner\neg\delta\urcorner \geqslant 0.2)).$$

An agent's credences is a function assigning to each sentence (in $\mathcal{A}$) a real number.

Definition 6.2.3. The agent's possible credence functions are given by $\mathsf{Creds}_{\mathsf{Sent}_\mathsf{P}}$. These are all functions from Sent_P to $[0,1]$.

Her possible credence functions, restricted to an agenda $\mathcal{A}$, are given by $\mathsf{Creds}_{\mathcal{A}}$, which consists of all functions from $\mathcal{A}$ to $[0,1]$.

Here we are making the assumption that an agent always assigns credences in the unit interval[3]. That is an assumption we will consider dropping in Section 7.2.3. It will turn out that this assumption is substantial if we accept Caie's suggested modification of the accuracy considerations as it will then not be justified by accuracy considerations.

As discussed in the introduction, we are here assuming that if the agent's credences are c then c is the correct interpretation of P and the agent is herself aware of this.[4] We will also only consider agendas where a choice of an agent's credences (in sentences in the agenda) determine all truths of sentences in that agenda. These will be called self-ref agendas.

Definition 6.2.4. $\mathcal{A} \subseteq \mathsf{Sent}_\mathsf{P}$ is a *self-ref agenda* if:

For all $c, c' \in \mathsf{Creds}_{\mathsf{Sent}_\mathsf{P}}$, and $\mathcal{M}_c, \mathcal{M}_{c'} \in \mathsf{Mod}^{\mathsf{PA,ROCF}}_{\mathcal{L}_\mathsf{P}}$, such that for any $\varphi \in \mathcal{A}$,

$$\mathcal{M}_c \models \mathsf{P}\ulcorner\varphi\urcorner = r \iff r = c(\varphi)$$
$$\text{and } \mathcal{M}_{c'} \models \mathsf{P}\ulcorner\varphi\urcorner = r \iff r = c'(\varphi)$$

we have:

$$\forall\varphi \in \mathcal{A}, (c(\varphi) = c'(\varphi)) \implies \forall\varphi \in \mathcal{A}, (\llbracket\varphi\rrbracket_{\mathcal{M}_c} = \llbracket\varphi\rrbracket_{\mathcal{M}_{c'}})$$

There are two important features of a self-ref agenda. The first is that they don't contain any sentences whose truth is determined by empirical matters,

---

[2]We could take it to be in some other sense necessary by altering $\mathsf{Mod}^{\mathsf{PA,ROCF}}_{\mathcal{L}_\mathsf{P}}$ to some other set of models which satisfies the biconditional. That then may help us to model situations like Alice's promotion by just imposing that all considered situations, or worlds, satisfy the relevant biconditional $Promotion \leftrightarrow \neg\mathsf{P}^{\mathrm{Alice}}\ulcorner Promotion\urcorner \geqslant {}^1\!/_2$.

[3]$= [0,1] = \{r \in \mathbb{R} \mid 0 \leqslant r \leqslant 1\}$.

[4]This is basically assuming that the probabilistic modal structure representing her is introspective. Due to the challenges with introspection and probabilism these are the cases in which we'll have a problem. However, in this thesis we have been focusing on not simply rejecting assumptions like introspection but to see how to deal with them in the self-referential framework, and this methodology we will continue with here.

the truth of all sentences can only depend on the interpretation of P. Secondly they have a fullness component: if a sentence is in the agenda and it refers to the probability of some other sentences, then these other sentences must also be in the agenda.

In Chapters 7 and 8 we will only consider self-ref agendas. We are therefore uninterested in the agent's uncertainty about anything whose truth isn't dependent on her own credences, e.g. we are not interested in her degree of belief in *Heads*.

For example the following are self-ref agendas

Example 6.2.5. • Consider $\pi \leftrightarrow \neg\mathsf{P}\ulcorner\pi\urcorner \geqslant 1/2$. $\{\pi\}$ is itself a self-ref agenda, and so is $\{\pi, \neg\pi\}$, and also $\{\pi, \neg\pi, \top \vee \pi, \pi \wedge \pi, \bot\}$. However, $\{\pi, Heads\}$ is not.

- Let $\delta \leftrightarrow (\mathsf{P}\ulcorner\delta\urcorner < 1/2 \vee \mathsf{P}\ulcorner\neg\delta\urcorner \geqslant 1/2))$. Then a self-ref agenda including δ would also need to include $\neg\delta$. In fact $\{\delta, \neg\delta\}$ is a self-ref agenda.

- Let $\delta \leftrightarrow \mathsf{P}\ulcorner\delta'\urcorner = 1$ and $\delta' \leftrightarrow \mathsf{P}\ulcorner\delta\urcorner > 0$. Then $\{\delta, \delta'\}$ is a self-ref agenda.

- Let $\delta \leftrightarrow \mathsf{P}\ulcorner\mathsf{P}\ulcorner\delta\urcorner > 1/2\urcorner = 1$. Then a self-ref agenda including δ would also need to include $\mathsf{P}\ulcorner\delta\urcorner > 1/2$. In fact $\{\delta, \mathsf{P}\ulcorner\delta\urcorner > 1/2\}$ is a self-ref agenda.

An agent's credences are judged with respect to something. It is quite common in talking about rational requirements to use the term "world" to refer to the matters which make sentences true and false. And it is such worlds with respect to which the agent's credences are judged. In Part I of this thesis we were using "worlds", or "w" to refer to objects in our probabilistic modal structures. In this part of the thesis we will use sans-serif "w" to give truth values of the relevant sentences in $\mathcal{A}$. This is essentially some model, $\mathcal{M}$, restricted just to $\mathcal{A}$.

Definition 6.2.6. Let $\mathcal{A}$ be any agenda. For $\mathcal{M} \in \mathsf{Mod}_{\mathcal{L}_\mathsf{P}}$, define $\mathcal{M}\restriction_{\mathcal{A}}$ be a function from $\mathcal{A}$ to $\{0, 1\}$ by

$$\mathcal{M}\restriction_{\mathcal{A}}(\varphi) := [\![\varphi]\!]_{\mathcal{M}} = \begin{cases} 1 & \mathcal{M} \models \varphi \\ 0 & \text{otherwise} \end{cases}.$$

Define:
$$\mathsf{Worlds}_{\mathcal{A}} := \{\mathcal{M}\restriction_{\mathcal{A}} \mid \mathcal{M} \in \mathsf{Mod}_{\mathcal{L}_\mathsf{P}}^{\mathsf{PA,ROCF}}\}.$$

We use w as a metavariable for a member of $\mathsf{Worlds}_{\mathcal{A}}$.

Definition 6.2.7. If $\mathcal{A}$ is a self-ref agenda where $\mathsf{Worlds}_{\mathcal{A}}$ has just two members, one where δ is true, and one where δ is false, we may also use the notation w_δ and $\mathsf{w}_{\neg\delta}$ to refer to these. In that case we say that $\mathcal{A}$ is a δ-*agenda*.

Equivalently, $\mathcal{A}$ is a δ-agenda iff it is a self-ref agenda and for every $\varphi \in \mathcal{A}$

$$\mathsf{PA} \vdash \begin{array}{l} ((\delta \to \varphi) \vee (\delta \to \neg\varphi)) \\ \wedge ((\neg\delta \to \varphi) \vee (\neg\delta \to \neg\varphi)) \end{array}$$

For example all the examples of self-ref agendas involving π above were π-agendas. For any δ, if $\{\delta, \neg\delta\}$ is a self-ref agenda then it is also a δ-agenda. But for $\delta \leftrightarrow \mathsf{P}^\ulcorner \mathsf{P}^\ulcorner \delta^\urcorner \geqslant 1/2^\urcorner = 1$, there are no δ-agendas.[5]

Note that all δ-agendas are self-ref agendas by definition.

As mentioned in the introduction, Caie's proposal is that one should not judge an agent's credal state with respect to all worlds, but only those that are consistent with the agent having the considered credences. In the case of self-ref agendas, a choice of credal state *determines* the truths of sentences and so the world at stake. We therefore define w_c to refer to this the world that would be actual if the agent had the credences c.

Definition 6.2.8. Let $\mathcal{A}$ be a self-ref agenda. Suppose $c \in \mathsf{Creds}_\mathcal{A}$. Define $\mathsf{w}_c = \mathcal{M}_c \restriction_\mathcal{A}$, where $\mathcal{M}_c \in \mathsf{Mod}^{\mathsf{PA,ROCF}}_{\mathcal{L}_\mathsf{P}}$ is such that for any $\varphi \in \mathcal{A}$,[6]

$$\mathcal{M}_c \models \mathsf{P}^\ulcorner \varphi^\urcorner = r \iff r = c(\varphi)$$

If $\mathcal{A}$ is a δ-agenda, we have some $R \subseteq \mathsf{Creds}_\mathcal{A}$ such that

$$\delta \leftrightarrow {}`c \in R\text{'}.$$

This could be done by defining $R = \{c \mid \mathsf{w}_c = \mathsf{w}_\delta\}$.

For a δ-agenda we can represent this situation diagrammatically.

Example 6.2.9. The situation for π (where $\pi \leftrightarrow \neg\mathsf{P}^\ulcorner \pi^\urcorner \geqslant 1/2$) with $\mathcal{A} = \{\pi, \neg\pi\}$ can be represented as in Fig. 6.1. The agent's credence can be represented by some point in the square $[0,1]^2$. If the agent's credence is represented by a point in the shaded region then π. Otherwise π is false.

Note that this image is 2-dimensional because $\mathcal{A}$ just contains two sentences. Generally δ-agendas may contain arbitrarily many sentences then the corresponding "diagram" would be an n-dimensional one.

---

[5] A self-ref agenda must involve $\mathsf{P}^\ulcorner \delta^\urcorner \geqslant 1/2$. Consider $\mathcal{M}, \mathcal{M}' \in \mathsf{Mod}^{\mathsf{PA,ROCF}}_{\mathcal{L}_\mathsf{P}}$ given by the two interpretations of P, c and c' respectively:

$$c(\mathsf{P}^\ulcorner \delta^\urcorner \geqslant 1/2) = 1,\ c(\delta) = 1,$$
$$c'(\mathsf{P}^\ulcorner \delta^\urcorner \geqslant 1/2) = 1,\ c'(\delta) = 0.$$

And observe that

$$\mathcal{M} \models \delta \text{ since } \mathcal{M} \models \mathsf{P}^\ulcorner \mathsf{P}^\ulcorner \delta^\urcorner \geqslant 1/2^\urcorner = 1$$
$$\mathcal{M} \models \mathsf{P}^\ulcorner \delta^\urcorner \geqslant 1/2$$
$$\mathcal{M}' \models \delta \text{ since } \mathcal{M}' \models \mathsf{P}^\ulcorner \mathsf{P}^\ulcorner \delta^\urcorner \geqslant 1/2^\urcorner = 1$$
$$\mathcal{M}' \models \neg\mathsf{P}^\ulcorner \delta^\urcorner \geqslant 1/2$$

So

$$\mathsf{PA} \nvdash (\delta \to \mathsf{P}^\ulcorner \delta^\urcorner \geqslant 1/2) \vee (\delta \to \neg\mathsf{P}^\ulcorner \delta^\urcorner \geqslant 1/2)$$

[6] Note that the choice of $\mathcal{M}_c$ doesn't matter because we have assumed that $\mathcal{A}$ is a self-ref agenda.
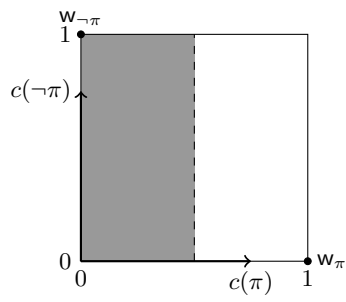
Figure 6.1: A diagram representing the agent's possible credences in π and $\neg\pi$ (restricted to those in $[0,1]^2$). The shaded area is the region where if the agent's credences are in that region then π is true.

Chapter 7

Accuracy

The accuracy argument assumes that the epistemic goal of an agent is to maximise the accuracy of their credences. This style of rational constraint was proposed in Joyce (1998).

Rationality Criterion 1 (Usual Accuracy Criterion). *An agent is irrational if she has a credal state b which is dominated in accuracy.*

I.e. $b \in \mathsf{Creds}_{\mathcal{A}}$ is irrational if there is some $c \in \mathsf{Creds}_{\mathcal{A}}$ such that for all $\mathsf{w} \in \mathsf{Worlds}_{\mathcal{A}}$,

$$\mathcal{I}(c, \mathsf{w}) < \mathcal{I}(b, \mathsf{w}).$$

There are in fact different variations of this criterion, but these won't in fact affect us much. A good overview can be found in Pettigrew (ms).

This requirement refers to some inaccuracy measure, or $\mathcal{I}$.

Definition 7.0.10. An *inaccuracy measure*, $\mathcal{I}$, is some function

$$\mathcal{I} : \bigcup_{\substack{\mathcal{A} \text{ is a} \\ \text{finite agenda}}} (\mathsf{Creds}_{\mathcal{A}} \times \mathsf{Worlds}_{\mathcal{A}}) \to [0, \infty].$$

Since we only work with finite agendas here, we do not need to assume that $\mathcal{I}$ is defined on infinite agendas. A common choice of an inaccuracy measure is the Brier score:

Definition 7.0.11. For $\mathcal{A} = \{\varphi_1, \ldots, \varphi_n\}$ the Brier score (BS) is such that:

$$\mathsf{BS}(c, \mathsf{w}) := \sum_{i=1}^{n} \frac{1}{n} \cdot (c(\varphi_i) - \mathsf{w}(\varphi_i))^2$$

A wide class of inaccuracy measures in fact lead to the same rationality requirement according to Usual Accuracy Criterion: that an agent should be probabilistic.

7.1 Caie's decision-theoretic understanding

7.1.1 The criterion

Caie (2013) argued that when one works with frameworks that have sentences whose truth depends on the probability they are assigned, one should only

consider the accuracy of an agent's credences at worlds where the agent has those credences. This would be the natural way of understanding the criterion in a decision theoretic manner and considering the utility of the act of coming to occupy some credal state.

Rationality Criterion 2 (Minimize Self-Inaccuracy). *Let $\mathcal{A}$ be a self-ref agenda. An agent should minimize*

$$\mathsf{SelfInacc}^{\mathcal{I}}(c) := \mathcal{I}(c, \mathsf{w}_c)$$

When $\mathcal{I}$ is clear from context we will drop the reference to it.

For $\mathcal{A}$ a δ-agenda, with $\delta \leftrightarrow \text{'} c \in R \text{'}$ this is:

$$\mathsf{SelfInacc}^{\mathcal{I}}(c) = \begin{cases} \mathcal{I}(c, \mathsf{w}_\delta) & c \in R \\ \mathcal{I}(c, \mathsf{w}_{\neg\delta}) & c \notin R \end{cases}$$

Consider the following example as discussed in Caie (2013).

Example 7.1.1. Consider $\pi \leftrightarrow \neg\mathsf{P}^{\ulcorner}\pi^{\urcorner} \geqslant 1/2$ and $\mathcal{A} = \{\pi, \neg\pi\}$. As before we can represent π diagrammatically. But we can also include in this diagram some additional information about how the accuracy is measured.

If we consider some c in the grey region, then if the agent has those credences then w_π would be actual, i.e. $\mathsf{w}_c = \mathsf{w}_\pi$. For such points we should consider $\mathcal{I}(c, \mathsf{w}_\pi)$. If $\mathcal{I}$ is truth-directed this can be considered as some (generalised) notion of distance from that point to w_π. And similarly if we pick some point in the unshaded region to be the agent's credences, then the inaccuracy is measured from that point to $\mathsf{w}_{\neg\pi}$.

For inaccuracy measured by the Brier score, the minimum such distance is obtained at the point $\langle 1/2, 1 \rangle$, i.e. credences where $c(\pi) = 1/2$ and $c(\neg\pi) = 1$.
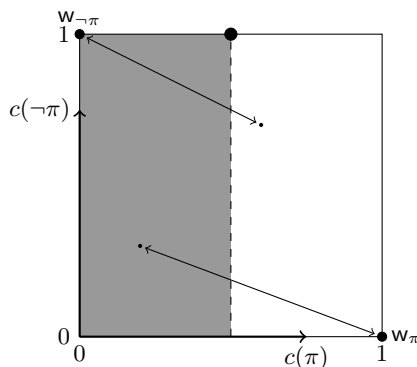
Figure 7.1: Measuring the accuracy for π

As Caie discusses, this criterion leads to the rejection of probabilism, at least when the Brier score is used, since it leads to a rational requirement to have credences $c(\pi) = 1/2$ and $c(\neg\pi) = 1$, a credal state that is not probabilistic.

We will now present some results which allow us to generally determine when a credal state is rationally required according to this criterion. This will allow us to then show that the criterion is very flexible and some undesirable consequences of this flexibility.

7.1.2 When b minimizes SelfInacc

Definition 7.1.2. Fix $\mathcal{A}$ a self-ref agenda. Let $b \in \mathsf{Creds}_{\mathcal{A}}$ and $\mathsf{w} \in \mathsf{Worlds}_{\mathcal{A}}$. Define

$$\mathsf{MoreAcc}_b^{\mathcal{I}}(\mathsf{w}) := \{c \in \mathsf{Creds}_{\mathcal{A}} \mid \mathcal{I}(c, \mathsf{w}) < \mathsf{SelfInacc}(b)\}$$
$$\mathsf{wMoreAcc}_b^{\mathcal{I}}(\mathsf{w}) := \{c \in \mathsf{Creds}_{\mathcal{A}} \mid \mathcal{I}(c, \mathsf{w}) \leqslant \mathsf{SelfInacc}(b)\}$$

and for $r \in \mathbb{R}$ define:

$$\mathsf{MoreAcc}_r^{\mathcal{I}}(\mathsf{w}) := \{c \in \mathsf{Creds}_{\mathcal{A}} \mid \mathcal{I}(c, \mathsf{w}) < r\}$$
$$\mathsf{wMoreAcc}_r^{\mathcal{I}}(\mathsf{w}) := \{c \in \mathsf{Creds}_{\mathcal{A}} \mid \mathcal{I}(c, \mathsf{w}) \leqslant r\}$$

When $\mathcal{I}$ is clear from context we will drop the reference to it.

Later we will consider $\mathcal{I}$s which satisfy Extensionality, and in such a case we can also use the versions with r without first picking $\mathcal{A}$.

For an example of $\mathsf{MoreAcc}_b^{\mathsf{BS}}(\mathsf{w})$ consider Fig. 7.2.
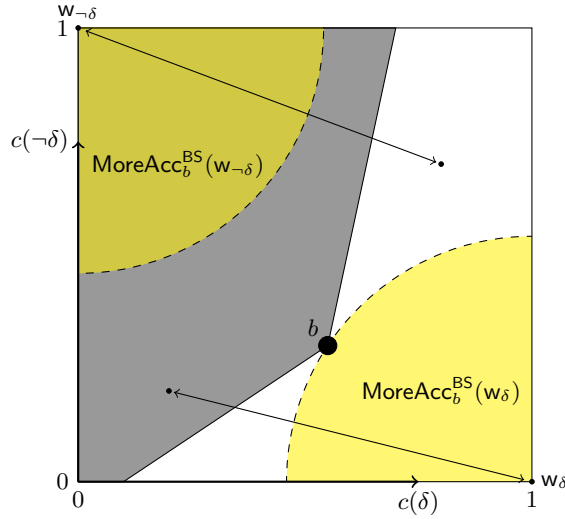
Figure 7.2: An example of $\mathsf{MoreAcc}_b^{\mathsf{BS}}(\mathsf{w})$

$\mathsf{wMoreAcc}_b^{\mathsf{BS}}(\mathsf{w})$ would also include the edge.

The following is a trivial observation:

Proposition 7.1.3. *Let $\mathcal{A}$ be a self-ref agenda.*

$$\mathsf{SelfInacc}(c) \geqslant \mathsf{SelfInacc}(b) \iff c \notin \mathsf{MoreAcc}_b(\mathsf{w}_c)$$
$$\mathsf{SelfInacc}(c) > \mathsf{SelfInacc}(b) \iff c \notin \mathsf{wMoreAcc}_b(\mathsf{w}_c)$$

Which immediately implies:

Proposition 7.1.4. *Let $\mathcal{A}$ be a self-ref agenda.*
$\mathsf{SelfInacc}$ is minimised (possibly non-uniquely) at b iff

$$\textit{for all } c \in \mathsf{Creds}_{\mathcal{A}}, \ c \notin \mathsf{MoreAcc}_b(\mathsf{w}_c)$$

SelfInacc *is minimised uniquely at b iff*

$$for\ all\ c \in \mathsf{Creds}_{\mathcal{A}} \setminus \{b\},\ c \notin \mathsf{wMoreAcc}_b(\mathsf{w}_c)$$

In the case where $\mathcal{A}$ is a δ-agenda we have a further characterisation of when a credal state minimizes SelfInacc:

Proposition 7.1.5. *Let $\mathcal{A}$ be a δ-agenda, and let $R = \{c \in \mathsf{Creds}_{\mathcal{A}} \,|\, \mathsf{w}_c = \mathsf{w}_\delta\}$.[1]* SelfInacc *is minimised (possibly non-uniquely) at b iff*

$$R \cap \mathsf{MoreAcc}_b(\mathsf{w}_\delta) = \varnothing$$
$$and \quad R \supseteq \mathsf{MoreAcc}_b(\mathsf{w}_{\neg\delta}).$$

SelfInacc *is minimised uniquely at b iff*

$$R \cap \mathsf{wMoreAcc}_b(\mathsf{w}_\delta) \subseteq \{b\}$$
$$and \quad R \supseteq \mathsf{wMoreAcc}_b(\mathsf{w}_{\neg\delta}) \setminus \{b\}$$

To get an idea of why this is, fix the Brier score and consider the example in Fig. 7.2. In this diagram the point labelled b minimizes SelfInacc uniquely. This is because: for there to be a point in R that has smaller SelfInacc than b, it would have to lie in region labelled $\mathsf{MoreAcc}_b(\mathsf{w}_\delta)$, and one not in R would have to be in $\mathsf{MoreAcc}_b(\mathsf{w}_{\neg\delta})$. The points which have the same SelfInacc as b would either lie on the dotted edge of $\mathsf{MoreAcc}_b(\mathsf{w}_\delta)$ and be in R or lie on the dotted edge of $\mathsf{MoreAcc}_b(\mathsf{w}_{\neg\delta})$ and lie outside R. The general proof of the result just states this argument in a more general way.

Proof. By Proposition 7.1.4, b minimizes SelfInacc (possibly non-uniquely) iff

$$for\ all\ c \in \mathsf{Creds}_{\mathcal{A}}, \qquad \begin{array}{l} c \in R \text{ and } c \notin \mathsf{MoreAcc}_b(\mathsf{w}_\delta) \\ \text{or } c \notin R \text{ and } c \notin \mathsf{MoreAcc}_b(\mathsf{w}_{\neg\delta}) \end{array}$$

This holds iff

$$R \cap \mathsf{MoreAcc}_b(\mathsf{w}_\delta) = \varnothing$$
$$and \quad R \supseteq \mathsf{MoreAcc}_b(\mathsf{w}_{\neg\delta}).$$

Also by Proposition 7.1.4, b minimizes SelfInacc uniquely iff

$$for\ all\ c \in \mathsf{Creds}_{\mathcal{A}} \text{ with } c \neq b, \qquad \begin{array}{l} c \in R \text{ and } c \notin \mathsf{wMoreAcc}_b(\mathsf{w}_\delta) \\ \text{or } c \notin R \text{ and } c \notin \mathsf{wMoreAcc}_b(\mathsf{w}_{\neg\delta}) \end{array}$$

This holds iff

$$R \cap \mathsf{wMoreAcc}_b(\mathsf{w}_\delta) \subseteq \{b\}$$
$$and \quad R \supseteq \mathsf{wMoreAcc}_b(\mathsf{w}_{\neg\delta}) \setminus \{b\} \qquad \square$$

As a direct consequence of this we have the following:

---

[1]So we have $\delta \leftrightarrow$ '$c \in R$'.

Proposition 7.1.6. *If* SelfInacc *is minimised (possibly non-uniquely) at b, then*

$$\mathsf{MoreAcc}_b(\mathsf{w}_\delta) \cap \mathsf{MoreAcc}_b(\mathsf{w}_{\neg\delta}) = \varnothing$$

If SelfInacc *is minimized uniquely at b then*

$$\mathsf{wMoreAcc}_b(\mathsf{w}_\delta) \cap \mathsf{wMoreAcc}_b(\mathsf{w}_{\neg\delta}) \subseteq \{b\}$$

We will also show that if we impose an extra constraint on $\mathcal{I}$ we can get something like a converse of this result. That will allow us to show that the Minimize Self-Inaccuracy is in fact a very flexible criterion.

7.1.3 The flexibility

For this result, we need to make some minimal assumptions on the inaccuracy measure $\mathcal{I}$.

Definition 7.1.7. $\mathcal{I}$ satisfies Extensionality iff:
   Suppose $\mathcal{A} = \{\varphi_1, \ldots, \varphi_n\}$ and $\mathcal{A}' = \{\varphi_1', \ldots, \varphi_n'\}$ are self-ref agendas. Suppose $c \in \mathsf{Creds}_\mathcal{A}$ and $c' \in \mathsf{Creds}_{\mathcal{A}'}$ are such that

$$\text{for all } i, \ c(\varphi_i) = c'(\varphi_i')$$

And w and w' are such that

$$\text{for all } i, \ \mathsf{w}(\varphi_i) = \mathsf{w}'(\varphi_i')$$

Then

$$\mathcal{I}(c, \mathsf{w}) = \mathcal{I}(c', \mathsf{w}').$$

This says that $\mathcal{I}$ does not depend on the particular sentence, it only depends on the multiset:[2]

$$\{\langle c(\varphi), \mathsf{w}(\varphi)\rangle \mid \varphi \in \mathcal{A}\}.$$

Proposition 7.1.8. *If* $\mathcal{I}$ *satisfies Extensionality then there is some* $d : \bigcup_n (\mathbb{R}^n \times \{0,1\}^n) \to \mathbb{R}$ *such that for* $c \in \mathsf{Creds}_\mathcal{A}, \mathsf{w} \in \mathsf{Worlds}_\mathcal{A},$[3]

$$\mathcal{I}(c, \mathsf{w}) = d_\mathcal{I}(\langle c(\varphi_1), \ldots, c(\varphi_n)\rangle, \langle \mathsf{w}(\varphi_1), \ldots, \mathsf{w}(\varphi_n)\rangle).$$

It might, however, fail to satisfy Normality, which says that the inaccuracy score is a function of $|c(\varphi) - \mathsf{w}(\varphi)|$, because an inaccuracy function satisfying Extensionality may still care about closeness to truth more than it cares about closeness to falsity. However most natural functions satisfying Extensionality will also satisfy Normality.

Once we have assumed that $\mathcal{I}$ satisfies Extensionality we can consider the inaccuracy of a credal state just by considering it as a point in $\mathbb{R}^2$ and forgetting which agenda it is associated with.

---

[2] Which is unordered but may contain repetitions.
[3] In fact it also cannot depend on the ordering simply because we assumed that $\mathcal{A}$ was unordered.

Definition 7.1.9. Suppose $\mathcal{I}$ satisfies Extensionality. Suppose $\mathbf{x} = \langle x_1, \ldots, x_n \rangle \in \mathbb{R}^n$. Consider a δ-agenda $\mathcal{A} = \{\varphi_1, \ldots, \varphi_n\}$.[4]

Then we can let

$$\mathcal{I}(\mathbf{x}, \mathsf{w}_n^{\mathsf{pos}})$$

be equal to $\mathcal{I}(c, \mathsf{w}_\delta)$ where $c(\varphi_i) = x_i$. Similarly

$$\mathcal{I}(\mathbf{x}, \mathsf{w}_n^{\mathsf{neg}})$$

be equal to $\mathcal{I}(c, \mathsf{w}_{\neg\delta})$ for such c.

Proposition 7.1.10. *Suppose $\mathcal{I}$ satisfies Extensionality. For any δ-agenda, $\mathcal{A} = \{\varphi_1, \ldots, \varphi_n\}$,*

$$\mathcal{I}(c, \mathsf{w}_\delta) = \mathcal{I}(\mathbf{x}_c, \mathsf{w}_n^{\mathsf{pos}})$$
$$\mathcal{I}(c, \mathsf{w}_{\neg\delta}) = \mathcal{I}(\mathbf{x}_c, \mathsf{w}_n^{\mathsf{neg}})$$

for $\mathbf{x}_c := \langle c(\varphi_1), \ldots, c(\varphi_n) \rangle$.

If we assume that enough regions are definable in the language and Extensionality is satisfied we will be able to show:

For any $\mathbf{x} \in \mathbb{R}^n$ that is "$\mathcal{I}$-close enough" to either $\mathsf{w}_n^{\mathsf{pos}}$ or $\mathsf{w}_n^{\mathsf{neg}}$, there is some δ and a δ-agenda $\mathcal{A}$ such that the agent should have credences $c(\varphi_i) = x_i$ according to Minimize Self-Inaccuracy.

What we will mean by $\mathcal{I}$-close enough to $\mathsf{w}_n^{\mathsf{pos}}$ is that there is nothing that is closer to both $\mathsf{w}_n^{\mathsf{pos}}$ and $\mathsf{w}_n^{\mathsf{neg}}$ than $\mathbf{x}$ is to $\mathsf{w}_n^{\mathsf{pos}}$.

As an example of this "$\mathcal{I}$-close enough" we present the examples for BS, AbsValDist and ℓ^∞ in Fig. 7.3. We include the examples for AbsValDist and ℓ^∞ as these will be inaccuracy scores corresponding to the Dutch book criteria considered in Chapter 8, see Section 8.5.

For the shaded region in the diagrams in that figure we can find some sentence where the inaccuracy is minimized at that point.

This theorem has a very strong assumption, namely that many regions are definable. This will not be satisfied in $\mathcal{L}_\mathsf{P}$, but will in some expanded language.[5] In the case where $\mathcal{I}$ is continuous we can in fact find regions definable in $\mathcal{L}_{\mathsf{P}_{\geqslant r}}$ (Theorem 7.A.1).

Theorem 7.1.11. *Suppose $\mathcal{I}$ satisfies Extensionality and for the $R \subseteq \mathbb{R}^n$ required in the proof, there is some δ_R and a δ_R-agenda $\mathcal{A}_R$ such that $\delta_R \leftrightarrow$ '$\langle c(\varphi) \rangle_{\varphi \in \mathcal{A}_R} \in R$'*
    Define

$$\mathsf{DistClosestWorld}(\mathbf{x}) := \min\{\mathcal{I}(\mathbf{x}, \mathsf{w}_n^{\mathsf{pos}}), \mathcal{I}(\mathbf{x}, \mathsf{w}_n^{\mathsf{neg}})\}.$$

Then for $n \geqslant 2$:

$$\mathsf{MoreAcc}_{\mathsf{DistClosestWorld}(\mathbf{x})}(\mathsf{w}_n^{\mathsf{pos}}) \cap \mathsf{MoreAcc}_{\mathsf{DistClosestWorld}(\mathbf{x})}(\mathsf{w}_n^{\mathsf{neg}}) = \varnothing$$

---

[4]One could take $\{\pi, \neg\pi, \pi \vee \pi, \pi \vee (\pi \vee \pi), \ldots, \overbrace{\pi \vee (\pi \vee (\ldots(\pi \vee (\pi \vee \pi)\ldots))}^{n-2}\}$.

[5]Though this will involve some alterations to the framework as the language is then uncountable so we cannot code sentences up in arithmetic but will instead have to use some alternative syntax theory.
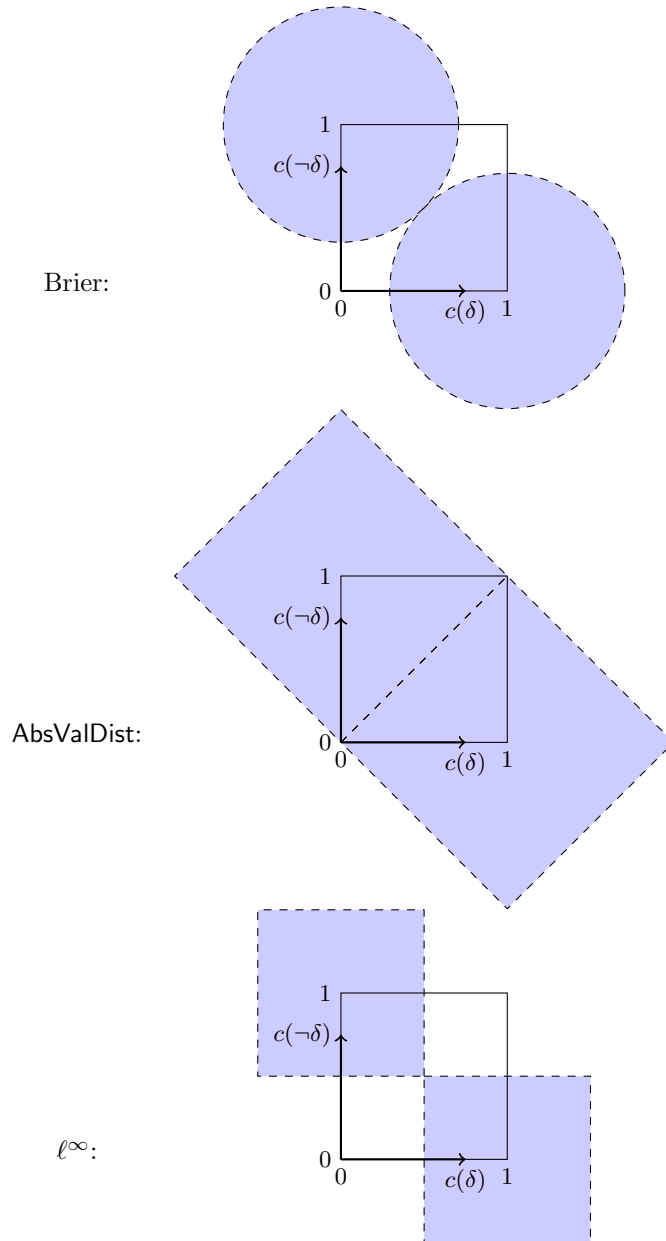
Figure 7.3: For any point in the shaded region, there is some sentence such that the agent is rationally required to have those credences. In this example we are dropping the assumption that credences lie in $[0, 1]$.

iff there is some δ-agenda $\{\varphi_1, \ldots, \varphi_n\}$[6] *such that* SelfInacc *is minimized (possibly non-uniquely) at* $b(\varphi_i) = x_i$.

And

$$\mathsf{wMoreAcc}_{\mathsf{DistClosestWorld}(\mathbf{x})}(\mathsf{w}_n^{\mathsf{pos}}) \cap \mathsf{wMoreAcc}_{\mathsf{DistClosestWorld}(\mathbf{x})}(\mathsf{w}_n^{\mathsf{neg}}) \subseteq \{\mathbf{x}\}$$

iff there is some δ-agenda $\{\varphi_1, \ldots, \varphi_n\}$ *such that* SelfInacc *is minimized uniquely at* $b(\varphi_i) = x_i$.

Proof. The right-to-left of these "iff" follow directly from Proposition 7.1.6, so we just need to show the left-to-right direction. We first consider the unique minimization.

Suppose

$$\mathsf{wMoreAcc}_{\mathsf{DistClosestWorld}(\mathbf{x})}(\mathsf{w}_n^{\mathsf{pos}}) \cap \mathsf{wMoreAcc}_{\mathsf{DistClosestWorld}(\mathbf{x})}(\mathsf{w}_n^{\mathsf{neg}}) \subseteq \{\mathbf{x}\}.$$

If $\mathcal{I}(\mathbf{x}, \mathsf{w}_n^{\mathsf{pos}}) \leqslant \mathcal{I}(\mathbf{x}, \mathsf{w}_n^{\mathsf{neg}})$, we can let

$$R := \{\mathbf{x}\} \cup (\mathbb{R}^n \setminus \mathsf{wMoreAcc}_{\mathsf{DistClosestWorld}(\mathbf{x})}(\mathsf{w}_n^{\mathsf{pos}})).$$

If $\mathcal{I}(\mathbf{x}, \mathsf{w}_n^{\mathsf{pos}}) \geqslant \mathcal{I}(\mathbf{x}, \mathsf{w}_n^{\mathsf{neg}})$, we can let

$$R := \mathsf{wMoreAcc}_{\mathsf{DistClosestWorld}(\mathbf{x})}(\mathsf{w}_n^{\mathsf{neg}}) \setminus \{\mathbf{x}\}.$$

Using Proposition 7.1.5 we can see that $\mathbf{x}$ will then uniquely minimize SelfInacc. This is displayed in Fig. 7.4.
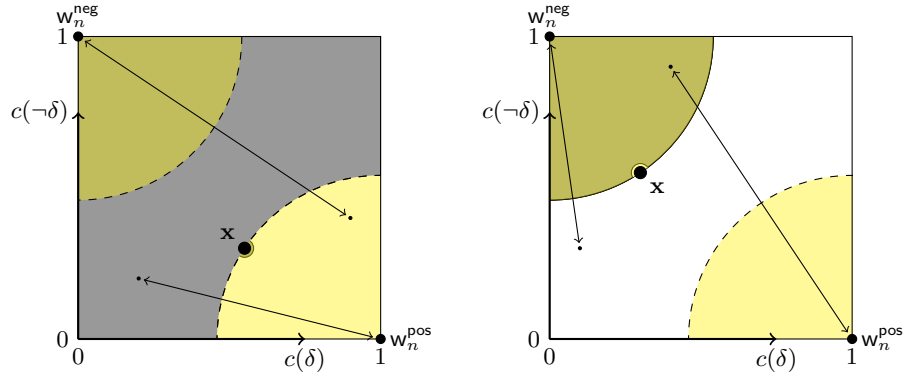
Figure 7.4: Defining R so $\mathbf{x}$ uniquely minimizes SelfInacc.

Suppose

$$\mathsf{MoreAcc}_{\mathsf{DistClosestWorld}(\mathbf{x})}(\mathsf{w}_n^{\mathsf{pos}}) \cap \mathsf{MoreAcc}_{\mathsf{DistClosestWorld}(\mathbf{x})}(\mathsf{w}_n^{\mathsf{neg}}) = \varnothing.$$

If $\mathcal{I}(\mathbf{x}, \mathsf{w}_n^{\mathsf{pos}}) \leqslant \mathcal{I}(\mathbf{x}, \mathsf{w}_n^{\mathsf{neg}})$, we can let $R := \mathbb{R}^n \setminus \mathsf{MoreAcc}_{\mathsf{DistClosestWorld}(\mathbf{x})}(\mathsf{w}_n^{\mathsf{pos}})$.

If $\mathcal{I}(\mathbf{x}, \mathsf{w}_n^{\mathsf{pos}}) \geqslant \mathcal{I}(\mathbf{x}, \mathsf{w}_n^{\mathsf{neg}})$, we can let $R := \mathsf{MoreAcc}_{\mathsf{DistClosestWorld}(\mathbf{x})}(\mathsf{w}_n^{\mathsf{neg}})$.

We then obtain possibly non-unique minimization at $\mathbf{x}$ using Proposition 7.1.5.

For a case where we have non-unique minimization see Fig. 7.5 □

---

[6] One could take $\{\delta, \neg\delta, \delta \vee \delta, \delta \vee (\delta \vee \delta), \ldots, \overbrace{\delta \vee (\delta \vee (\ldots (\delta \vee (\delta \vee \delta) \ldots))}^{n-2}\}$.
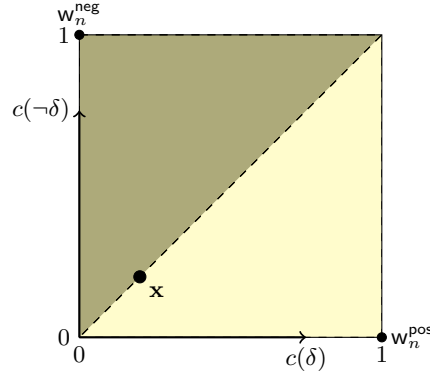
Figure 7.5: Defining R so $\mathbf{x}$ non-uniquely minimizes SelfInacc.

Without further assumptions on $\mathcal{I}$, this theorem is still quite limited because it might be that there is no $\mathbf{x}$ which is close enough to either of $\mathsf{w}_n^{\mathsf{pos}}$ or $\mathsf{w}_n^{\mathsf{neg}}$.

Example 7.1.12. Let $\mathcal{I}(c, \mathsf{w}) := 2$ for all c and w. Here $\mathsf{SelfInacc}(c) = 2$ for any c and δ.

Note that the theorem did not assume TruthDirectedness of $\mathcal{I}$.

7.2 Consequences of the flexibility

In this section we put forward some consequences of the flexibility of this rational requirement. We will later suggest that this gives us more reason to reject the rational requirement proposed. To be able to pick particular examples of the consequences of this criterion we will assume some additional things about the inaccuracy measure. However similar results will hold under weaker assumptions; for results using other assumptions Proposition 7.1.5 and Theorem 7.A.1 can be used.

Definition 7.2.1. An inaccuracy measure $\mathcal{I}$ satisfies TruthDirectedness iff for all $c, b \in \mathsf{Creds}_{\mathcal{A}}$:

- For all $\varphi \in \mathcal{A}$, either:

    - $\mathsf{w}(\varphi) \leqslant c(\varphi) \leqslant b(\varphi)$, or
    - $\mathsf{w}(\varphi) \geqslant c(\varphi) \geqslant b(\varphi)$

  and

- For some $\varphi \in \mathcal{A}$, either:

    - $\mathsf{w}(\varphi) \leqslant c(\varphi) < b(\varphi)$, or
    - $\mathsf{w}(\varphi) \geqslant c(\varphi) > b(\varphi)$

Then

$$\mathcal{I}(c, \mathsf{w}) < \mathcal{I}(b, \mathsf{w}).$$

175

It is only when $\mathcal{I}$ satisfies TruthDirectedness[7] that we can consider it as closeness to w.

In order to be able to pick concrete examples we will impose a further constraint on $\mathcal{I}$. This is Normality.

Definition 7.2.2. $\mathcal{I}$ satisfies Normality iff whenever $\mathcal{A} = \{\varphi_1, \ldots, \varphi_n\}$ and $\mathcal{A}' = \{\varphi'_1, \ldots, \varphi'_n\}$ are self-ref agendas and $c \in \mathsf{Creds}_{\mathcal{A}}$, $c' \in \mathsf{Creds}_{\mathcal{A}'}$, $\mathsf{w} \in \mathsf{Worlds}_{\mathcal{A}}$ and $\mathsf{w}' \in \mathsf{Worlds}_{\mathcal{A}'}$ are such that

$$\text{for all } i, \ |c(\varphi_i) - \mathsf{w}(\varphi_i)| = |c'(\varphi'_i) - \mathsf{w}'(\varphi'_i)|,$$

then

$$\mathcal{I}(c, \mathsf{w}) = \mathcal{I}(c', \mathsf{w}').$$

I.e. $\mathcal{I}$ is a function of the multiset

$$\{|c(\varphi) - \mathsf{w}(\varphi)| \mid \varphi \in \mathcal{A}\}.$$

Also note that Normality implies Extensionality.

7.2.1 Rejecting probabilism

Example 7.2.3 (Caie, unpublished result)**.** Consider $\mathcal{A} := \{\pi, \neg\pi\}$ where $\pi \leftrightarrow \neg\mathsf{P}^{\ulcorner}\pi^{\urcorner} \geqslant 1/2$.

Suppose $\mathcal{I}$ satisfies Normality and TruthDirectedness. Then b with $b(\pi) = 1/2$, $b(\neg\pi) = 1$ minimizes SelfInacc.

See Fig. 7.6
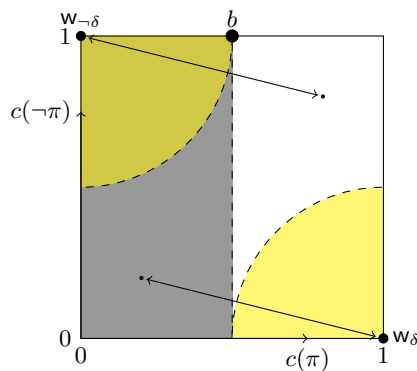
Figure 7.6: Rejection of probabilism

This generalises Example 7.1.1 to work for a class of inaccuracy scores instead of just the Brier score. The argument for it, though, is exactly the same.

---

[7]Or at least its weaker form:

- For all $\varphi \in \mathcal{A}$, either:
  - $\mathsf{w}(\varphi) \leqslant c(\varphi) \leqslant b(\varphi)$, or
  - $\mathsf{w}(\varphi) \geqslant c(\varphi) \geqslant b(\varphi)$

Then

$$\mathcal{I}(c, \mathsf{w}) \leqslant \mathcal{I}(b, \mathsf{w}).$$

7.2.2 Failure of introspection

Consider an example very closely related to π. Instead of considering $\pi \leftrightarrow \neg P^\ulcorner\pi^\urcorner \geqslant 1/2$, consider

$$\delta \leftrightarrow P^\ulcorner\delta^\urcorner \leqslant 1/2.$$

This will show that Minimize Self-Inaccuracy is inconsistent with a principle saying that introspection is rationally permissible.

Example 7.2.4. Consider $\mathcal{A} := \{\delta, \neg\delta\}$ where $\delta \leftrightarrow \neg P^\ulcorner\delta^\urcorner > 1/2$.

Suppose $\mathcal{I}$ satisfies Normality and TruthDirectedness. Then b with $b(\delta) = 1/2$, $b(\neg\delta) = 0$ minimizes SelfInacc.

See Fig. 7.7 This is a credal state where $b(\delta) \ngtr 1/2$, but $b(\neg P^\ulcorner\delta^\urcorner > 1/2) =$
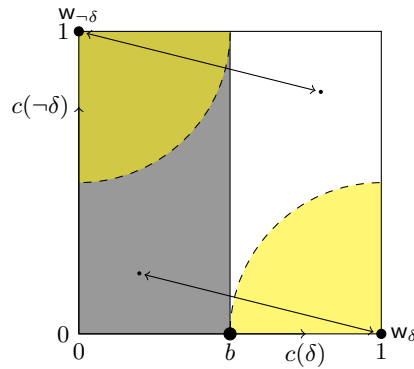
Figure 7.7: Rejection of introspection

$b(\neg\delta) = 0$. So if an agent is rational according to Minimize Self-Inaccuracy with a self-ref agenda containing δ, then she must very badly fail the principle:

$$\text{for all } \varphi \in \mathcal{A}, \quad c(\varphi) \ngtr 1/2 \implies c(\neg P^\ulcorner\delta^\urcorner > 1/2) = 1.$$

One can also show a similar result for the related *positive* introspection principle (now with $\leqslant$).

Example 7.2.5. Consider $\mathcal{A} := \{\delta\}$ where $\delta \leftrightarrow P^\ulcorner\delta^\urcorner \leqslant 1/2$.

Suppose $\mathcal{I}$ satisfies Normality and TruthDirectedness. Then b with $b(\delta) = 1/2$ minimizes SelfInacc. As shown in Fig. 7.8. This is a credal state where $b(\delta) \leqslant 1/2$,
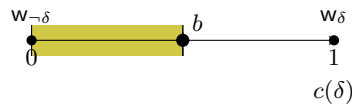
Figure 7.8: Another rejection of introspection

but $b(P^\ulcorner\delta^\urcorner \leqslant 1/2) = b(\delta) = 1/2$. So if an agent is rational according to Minimize Self-Inaccuracy with a self-ref agenda containing δ, then she cannot satisfy

$$\text{for all } \varphi \in \mathcal{A}, \quad c(\varphi) \leqslant 1/2 \implies c(P^\ulcorner\delta^\urcorner \leqslant 1/2) = 1.$$

This is a very interesting feature of this criterion. In Caie (2013), Caie supported two arguments for the possibility of rational probabilistic incoherence. The first was due to the inconsistency of probabilism with introspection, the second is due to Minimize Self-Inaccuracy. This result shows that although both might be reasons to reject probabilism, by dropping probabilism one has not removed inconsistencies between the desirable principles of introspection and Minimize Self-Inaccuracy.

7.2.3 Negative credences

So far we have generally been making the assumption that we only consider credences in the unit interval. It turns out that this was in fact an assumption that cannot be justified by Minimize Self-Inaccuracy, because if that assumption is dropped then sometimes negative credences minimize SelfInacc.

Example 7.2.6. For this example drop the assumption that an agent's credences take values in the unit interval.

Consider $\mathcal{A} := \{\delta\}$ where

$$\delta \leftrightarrow -1/4 < \mathsf{P}^\ulcorner\delta\urcorner < 1/2.$$

Fix $\mathcal{I}$ some inaccuracy measure satisfying Normality and TruthDirectedness. Then b with $b(\delta) = -1/4$ minimizes SelfInacc.

See Fig. 7.9
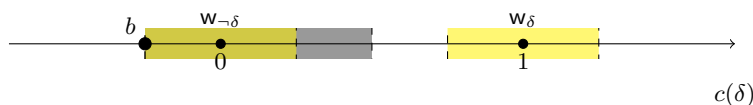
Figure 7.9: May require negative credences

Sometimes this assumption of having credence values in the unit interval *is* dropped, but this is generally only when one considers altering the numerical proxies of 0 and 1 for falsity and truth. What is interesting in this case is that we still keep the numerical representation of 0 for falsity and 1 for truth, and the distance measure is still as usual but we end up with credences outside of the unit interval.

Having a negative credence in such cases does seem odd, and one can rectify this, as we have done, by just building in the assumption that all credence functions assign numbers from the unit interval to sentences. However, this is now an assumption that cannot be justified by the accuracy considerations. So if one wishes to keep the rational requirement of minimizing SelfInacc, one should either have an alternative argument for that assumption and then build it in, or should admit that sometimes an agent must have negative credences.

7.2.4 Failure of simple logical omniscience

Again consider $\pi \leftrightarrow \neg\mathsf{P}^\ulcorner\pi\urcorner \geqslant 1/2$. But we now consider a different agenda. Let $\mathcal{A} := \{\pi, \pi \vee \top\}$. $\pi \leftrightarrow (\pi \vee \top)$ is a logical tautology. Moreover, it is a simple one which we would expect a rational agent to be able to recognise.

However we will show that Minimize Self-Inaccuracy requires that the agent has a different credence in π to $\pi \vee \top$. This is a very bad failure of the agent being probabilistic. It is a principle which is assumed in many cases. This means that the assumption that probabilities are assigned to propositions (or sets of possible worlds) is now an assumption that makes a difference. Of course one can build in that assumption, and only consider that c should minimize SelfInacc with respect to other members of $\mathsf{Creds}_\mathcal{A}$ that satisfy logical omniscience in an analogous way to assuming that credences take values in the unit intervals. But this is a bad feature of the proposed accuracy criterion that this cannot be justified by accuracy considerations.

Example 7.2.7. Consider $\mathcal{A} := \{\pi, \pi \vee \top\}$ where $\pi \leftrightarrow \neg\mathsf{P}\ulcorner\pi\urcorner > {}^1\!/_2.$[8]
 Fix $\mathcal{I}$ some inaccuracy measure satisfying Normality and TruthDirectedness. Then b with $b(\pi) = {}^1\!/_2$, $b(\pi \vee \top) = 0$ minimizes SelfInacc.
 See Fig. 7.10:
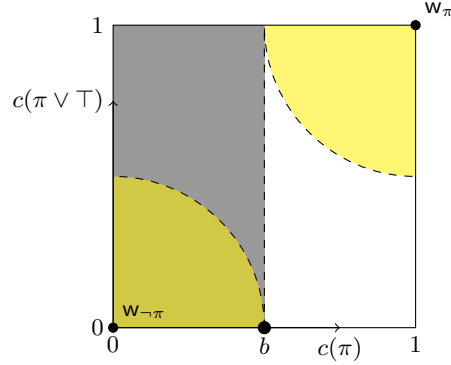
Figure 7.10: Rejection of logical omniscience

7.2.5 Dependence on the inaccuracy measure

By using Proposition 7.1.5 one can show that the rational requirement from Minimize Self-Inaccuracy is dependent on the inaccuracy measure that is chosen. For example, we consider the case of the Brier score vs the logarithmic score. This result was also presented in Campbell-Moore (2015b).

Theorem 7.2.8. *Consider:*

$$\delta \leftrightarrow (\mathsf{P}\ulcorner\delta\urcorner \leqslant 0.5 \vee (\mathsf{P}\ulcorner\delta\urcorner \leqslant 0.6 \wedge \mathsf{P}\ulcorner\neg\delta\urcorner \geqslant 0.2)).$$

Then

- $b = \langle 0, 0.5 \rangle$ *uniquely minimizes* $\mathsf{SelfInacc}^{\mathsf{AbsValDist}}$

- $b' = \langle 0.6, 0.2 \rangle$ *uniquely minimizes* $\mathsf{SelfInacc}^{\mathsf{BS}}$

- $b' = \langle 0.6, 0.2 \rangle$ *uniquely minimizes* $\mathsf{SelfInacc}^{\mathsf{LS}}$.
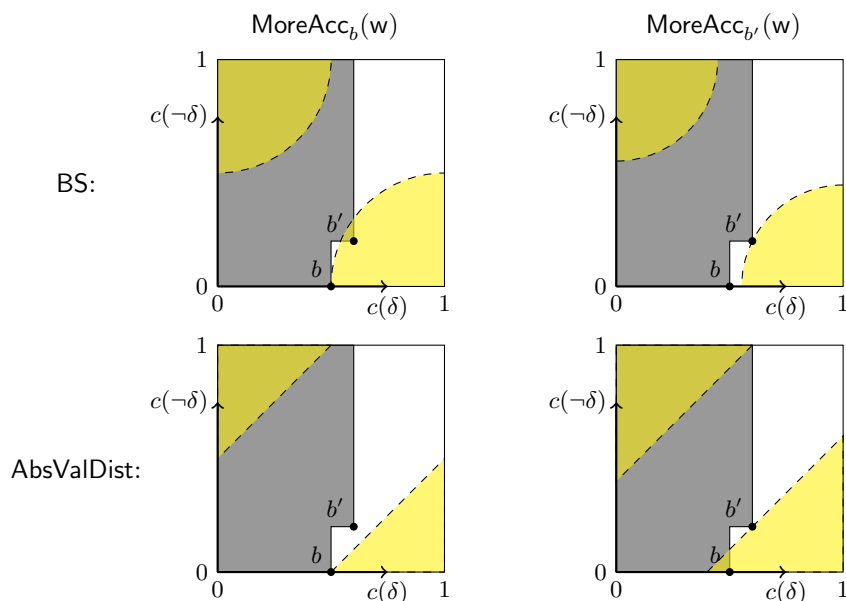
Figure 7.11: AbsValDist and BS lead to different rationality requirements

Proof. We will use Proposition 7.1.5 and for that purpose show the relevant balls in Fig. 7.11 for AbsValDist and BS.

Since all these inaccuracy measures are truth-directed and normal we can see that b and b' are the credal states that compete for being most accurate, and by calculating the values for these two credal states we can also see the result. This is how one can show the result for LS. □

This is not specific to these inaccuracy measure. We can find similar examples for most pairs of inaccuracy measures, at least when the measures aren't just scalings of one another at regions close to w_2^{pos} and w_2^{neg}. So even restricting to proper scoring rules doesn't lead to a determined class of rationally required credences.

Theorem 7.2.9. *Let $\mathcal{I}_0$ and $\mathcal{I}_1$ be inaccuracy measures satisfying Extensionality. And assume the language can define all relevant regions required in the proof.*

If there are some $q_0, q_1 \in \mathbb{R}$ such that all the following hold:

- q_0 *is $\mathcal{I}_0$-close-enough and q_1 is $\mathcal{I}_1$-close enough, meaning that none of balls around w_2^{pos} intersect any of the balls around w_2^{neg}. I.e. for all $i, j \in \{0, 1\}$,*

$$\mathsf{MoreAcc}_{q_i}^{\mathcal{I}_i}(w_2^{\mathsf{pos}}) \cap \mathsf{MoreAcc}_{q_j}^{\mathcal{I}_j}(w_2^{\mathsf{neg}}) = \varnothing.$$

- *And the balls aren't scalings of one another at these close enough distances to one of the w. I.e. for some $w \in \{w_2^{\mathsf{neg}}, w_2^{\mathsf{pos}}\}$*

$$\mathsf{MoreAcc}_{q_0}^{\mathcal{I}_0}(w) \not\subseteq \mathsf{MoreAcc}_{q_1}^{\mathcal{I}_1}(w)$$
$$\mathsf{MoreAcc}_{q_1}^{\mathcal{I}_1}(w) \not\subseteq \mathsf{MoreAcc}_{q_0}^{\mathcal{I}_0}(w)$$

---

[8] This result was observed by Benja Fallenstein (personal communication).

then there is some $R \subseteq \mathbb{R}^2$, δ with $\delta \leftrightarrow$ '$\langle c(\delta), c(\neg\delta) \rangle \in R$' and $\mathcal{A} = \{\delta, \neg\delta\}$, with some b_0 and $b_1 \in \mathsf{Creds}_{\mathcal{A}}$ such that:

- *$\mathsf{SelfInacc}^{\mathcal{I}_0}$ is minimized at b_0 and not at b_1, and*

- *$\mathsf{SelfInacc}^{\mathcal{I}_1}$ is minimized at b_1 and not at b_0*

So showing that the rational requirements of the criterion Minimize Self-Inaccuracy based on these two different inaccuracy measures are different.

Proof. Suppose we have such q_0, q_1. Without loss of generality suppose $\mathsf{w} = \mathsf{w}_2^{\mathsf{pos}}$. Then pick some

$$\langle x_0, y_0 \rangle \in \mathsf{MoreAcc}_{q_0}^{\mathcal{I}_0}(\mathsf{w}) \setminus \mathsf{MoreAcc}_{q_1}^{\mathcal{I}_1}(\mathsf{w})$$
$$\langle x_1, y_1 \rangle \in \mathsf{MoreAcc}_{q_1}^{\mathcal{I}_1}(\mathsf{w}) \setminus \mathsf{MoreAcc}_{q_0}^{\mathcal{I}_0}(\mathsf{w})$$

Now let

$$R := \mathbb{R}^2 \setminus (\mathsf{MoreAcc}_{\langle x_0, y_0 \rangle}^{\mathcal{I}_0}(\mathsf{w}_2^{\mathsf{pos}}) \cup \mathsf{MoreAcc}_{\langle x_1, y_1 \rangle}^{\mathcal{I}_1}(\mathsf{w}_2^{\mathsf{pos}}))$$

and $\delta \leftrightarrow$ ' $\langle c(\delta), c(\neg\delta) \rangle \in R$'. Now:

To show the facts about where $\mathsf{SelfInacc}$ is minimized we need to show:

1. $\mathsf{SelfInacc}^{\mathcal{I}_0}$ is minimized at $\langle x_0, y_0 \rangle$,

2. $\mathsf{SelfInacc}^{\mathcal{I}_0}$ is not minimized at $\langle x_1, y_1 \rangle$,

3. $\mathsf{SelfInacc}^{\mathcal{I}_1}$ is minimized at $\langle x_1, y_1 \rangle$,

4. $\mathsf{SelfInacc}^{\mathcal{I}_1}$ is not minimized at $\langle x_0, y_0 \rangle$,

Which we can do by using Proposition 7.1.5 so checking:

1. - $R \cap \mathsf{MoreAcc}_{\langle x_0, y_0 \rangle}^{\mathcal{I}_0}(\mathsf{w}_2^{\mathsf{pos}}) = \varnothing$, and
   - $R \supseteq \mathsf{MoreAcc}_{\langle x_0, y_0 \rangle}^{\mathcal{I}_0}(\mathsf{w}_2^{\mathsf{neg}})$,

2. - $R \cap \mathsf{MoreAcc}_{\langle x_1, y_1 \rangle}^{\mathcal{I}_0}(\mathsf{w}_2^{\mathsf{pos}}) \neq \varnothing$

3. - $R \cap \mathsf{MoreAcc}_{\langle x_1, y_1 \rangle}^{\mathcal{I}_1}(\mathsf{w}_2^{\mathsf{pos}}) = \varnothing$, and
   - $R \supseteq \mathsf{MoreAcc}_{\langle x_1, y_1 \rangle}^{\mathcal{I}_1}(\mathsf{w}_2^{\mathsf{neg}})$,

4. - $R \cap \mathsf{MoreAcc}_{\langle x_0, y_0 \rangle}^{\mathcal{I}_1}(\mathsf{w}_2^{\mathsf{pos}}) \neq \varnothing$.

These hold because of our choice of $\langle x_0, y_0 \rangle$ such that $\mathsf{SelfInacc}(\langle x_0, y_0 \rangle) \leqslant q_0$ and $\langle x_1, y_1 \rangle$ such that $\mathsf{SelfInacc}(\langle x_1, y_1 \rangle) \leqslant q_1$ and the assumptions about q_0 and q_1. $\qquad\square$

If one wishes to keep Minimize Self-Inaccuracy, one therefore needs to explain what rational constraints accuracy-dominance considerations lead to, and due to this dependence on the inaccuracy measure this is not as easy as it is for the usual accuracy constraint case.

There are at least four options.[9] Firstly, one could give arguments for one particular inaccuracy measure and argue that accuracy-dominance considerations require one to minimize inaccuracy with respect to that inaccuracy measure. Secondly, one could take a *subjectivist* approach and argue that for each agent and context there is some particular measure of inaccuracy which is appropriate. Thirdly, one could take a *supervaluationist* approach and argue that the notion of inaccuracy is vague and that any inaccuracy measure satisfying certain conditions is an appropriate precisification of it; to satisfy accuracy dominance considerations one would then have to minimise inaccuracy with respect to at least one appropriate inaccuracy measure. Lastly one could take an *epistemicist* approach and argue that although there is some particular inaccuracy measure which one should be minimising inaccuracy with respect to, we do not know which it is.[10]

This dependence on the inaccuracy measure has a further consequence. We will show in Section 8.5 that the Dutch book criterion proposed, Minimize Average-Unit-Guaranteed-Loss, is the same as Minimize Self-Inaccuracy with the inaccuracy measure AbsValDist (and that Minimize Maximum-Unit-Guaranteed-Loss is the same with the inaccuracy measure given by ℓ^∞). This result therefore shows that the Minimize Self-Inaccuracy criterion leads the agent to hold credences such that if she bets in accordance with those credences she does not minimize her possible overall guaranteed loss. I.e. that this accuracy criterion conflicts with the Dutch book criterion. In fact we will support neither this form of the accuracy or Dutch book criterion, but it is important to note that one cannot support both.

7.3 Accuracy criterion reconsidered

7.3.1 For introspective agents and self-ref agendas– the options

The results we have presented in Section 7.2 show that the rationality constraint proposed in Minimize Self-Inaccuracy, which says to minimize $\mathcal{I}(c, \mathsf{w}_c)$, leads to rational agents having very unwieldy credence functions. This might therefore give one a reason to reconsider the traditional accuracy criterion and suggest that it is in fact the correct way to apply accuracy considerations even when propositions like π are considered. One would still have to say how and why the traditional criterion does appropriately apply in such situations so our results haven't shown that this is the correct approach, just that the approach seems more tenable than it seemed before.

Responding to Greaves (2013), where Greaves considers cases that are related to π and other δ,[11] Konek and Levinstein (ms) and Carr (ms) suggest we

---

[9]This problem is very closely related to a problem for the traditional accuracy argument, the Bronfman objection, which is due to the fact that there is no credence function that dominates on every measure. This is discussed in (Pettigrew, 2011, section 6.2.2). Furthermore the ways of dealing with the two problems are similar and these options presented here parallel the options presented in Pettigrew's article.

[10]The disadvantage of this version of accuracy considerations is that an agent does not have the resources to know whether she satisfies the rational requirement or not.

[11]Greaves considers situations where the chance of something is related to its credence. So the version of *Promotion* in Greaves' paper has ch(*Promotion*) = $1 - c(Promotion)$.

should distinguish:

1. The inaccuracy of *the act of coming to occupy c*, which is given by $\mathsf{SelfInacc}(c) = \mathcal{I}(c, \mathsf{w}_c)$.

2. The inaccuracy of holding the *state c* from b's perspective, which is given by $\mathsf{Est}_b(\mathcal{I}(c, \cdot))$.

   The two options for how to understand this estimation are:[12]

   (a) $\mathsf{Exp}_b(\mathcal{I}(c, \cdot)) = \sum_{\mathsf{w} \in \mathsf{Worlds}_{\mathcal{A}}} b(\mathsf{w}) \cdot \mathcal{I}(c, \mathsf{w})$

   (b) $\mathcal{I}(c, \mathsf{w}_b)$.

Let's work through the example with π to see what these different possibilities would recommend.

Example 7.3.1. Suppose $b(\pi) = 1/3$, $b(\pi) = 2/3$ currently. The most accurate credal state from b's perspective is:[13]

1. $c_1(\pi) = 1/2$, $c_1(\neg\pi) = 1$.

2. (a) $c_{2a}(\pi) = 1/3$, $c_{2a}(\neg\pi) = 2/3$.

   (b) $c_{2b}(\pi) = 1$, $c_{2b}(\neg\pi) = 0$.
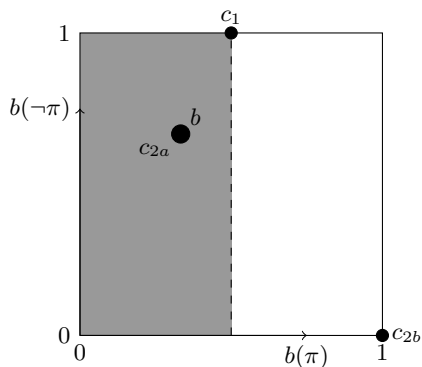
Figure 7.12: Alternative accuracy criteria

The accuracy criterion which we have been considering is to minimize the inaccuracy of the act of coming to occupy c. This is option 1.

Konek and Levinstein (ms) have argued that accuracy considerations should motivate us to try to minimize the inaccuracy of holding the state, not the inaccuracy of the act of coming to occupy the credal state.[14] This is because that has the correct direction of fit: the credences are trying to fit to the world instead of trying to force the world to fit to the credences. They argue as follows:

---

[12]Konek and Levinstein (ms) only considers the option which is analogous to 2b as an analysis of $\mathsf{Est}_b(\mathcal{I}(c, \cdot))$.

[13]For $\mathcal{I}$ satisfying Normality, TruthDirectedness, StrictPropriety.

[14]In fact they fix on option analogous to 2b, which we will also do, but the discussion of 2b vs 2a is an independent discussion.

> And credal states, as we have stressed, are better or worse (more or less valuable) to the extent to which they conform to the world by encoding an accurate picture of it.... But recall that credal states are *not* valuable in virtue of causally influencing the world so as to *make* themselves accurate. (Konek and Levinstein, ms, p. 19, their emphasis)

I agree with Konek and Levinstein that we should in fact consider a traditional accuracy criterion, which would be given by measuring estimated inaccuracy, namely either 2b or 2a. For further discussion of why that is an appropriate response, their paper should be consulted. Our results in Section 7.2 lend extra weight to rejecting inaccuracy as understood by 1.

Typically when dealing with the accuracy argument, $\mathsf{Est}_b(\mathcal{I}(c, \cdot))$ would be understood to be given by[15]

$$\mathsf{Est}_b(\mathcal{I}(c, \cdot)) = \mathsf{Exp}_b(\mathcal{I}(c, \cdot)) = \sum_{\mathsf{w} \in \mathsf{Worlds}_{\mathcal{A}}} b^*(\mathsf{w}) \cdot \mathcal{I}(c, \mathsf{w}).$$

This is the option in 2a.

Assuming that $\mathcal{I}$ satisfies a number of conditions, particularly StrictPropriety, we have that $\mathsf{Exp}_b(\mathcal{I}(c, \cdot))$ is minimized at b iff b is probabilistic. So this would not lead us to any rational constraint different to those in the usual accuracy arguments: an agent should just be probabilistic.

Usually higher order probabilities are not considered in accuracy arguments. When we do consider higher order probabilities this criterion doesn't seem to be the right thing to apply. In this criterion there is no information about what the internal P refers to; it could just as well be someone else's degrees of belief, or objective chance, or even some random non-probabilistic function. Consider the following example:

Example 7.3.2. Sophie is certain that the coin that has just been tossed has landed tails.

$$b(Heads) = 0$$
$$b(\neg Heads) = 1$$

She also has higher order beliefs. Let's suppose that she is perfectly introspective, which *should* mean that we get:

$$b(\mathsf{P}^{\mathrm{Sophie}} \ulcorner Heads \urcorner = 0) = 1,$$
$$b(\mathsf{P}^{\mathrm{Sophie}} \ulcorner \neg Heads \urcorner = 1) = 1.$$

But suppose Sophie's credences are not like that, but are instead such that:

$$b(\mathsf{P}^{\mathrm{Sophie}} \ulcorner Heads \urcorner = 1) = 1$$
$$b(\mathsf{P}^{\mathrm{Sophie}} \ulcorner \neg Heads \urcorner = 1) = 1$$

So Sophie's higher order beliefs do not cohere with her lower order ones even for non-problematic sentences like *Heads*. Even worse, Sophie is certain that $\mathsf{P}^{\mathrm{Sophie}}$ is non-probabilistic.[16]

---

[15]Following Pettigrew (ms) we take b^* to be some probabilistic function defined on $\mathfrak{B}(\mathcal{A})$, the smallest Boolean algebra extending $\mathcal{A}$.

[16]But note that b is itself probabilistic.

However, Sophie's beliefs, b, minimizes

$$\mathsf{Exp}_b(\mathcal{I}(c,\cdot)) = \sum_{\mathsf{w}\in\mathsf{Worlds}_{\mathcal{A}}} b^*(\mathsf{w}) \cdot \mathcal{I}(c,\mathsf{w}),$$

by virtue of $\mathcal{I}$ satisfying StrictProprietry and b being probabilistic.[17] Her beliefs are therefore *stable*.

We don't want to count Sophie as rational. We tried to assume that she was perfectly introspective but this was unable to play a role in the minimization of expected inaccuracy.

If we have supposed that an agent is introspective and we are considering self-ref agendas, she will be certain about which w is actual. And then the only thing that should matter is the distance to that w_b. This is embedded in the constraint 2b: the agent should determine the estimated inaccuracy by considering the inaccuracy of a credal state at the world the agent knows to be actual. This is different from looking at self-accuracy because the agent uses her *current* credences to determine what is actual instead of considering what would be actual *if* she were to adopt the credences she is considering.

In 2a the interpretation of P is left to vary, and b is instead just used to measure the possible varying interpretations of P. In 2b we use the credence state that she currently has, b, to *interpret* P, the symbol in our language as we are only considering the inaccuracy to w_b.

In Konek and Levinstein (ms), the authors in fact assume that the estimations are calculated in a way analogous to 2b instead of 2a. Following Greaves (2013), they formulate a principle called Deference to Chance which says that in some cases an agent *shouldn't* determine $\mathsf{Est}_b(\mathcal{I}(c,\cdot))$ by calculating the expectation.

> **DEFERENCE TO CHANCE** If an agent with credences c and evidence E is such that:
>
> $$\sum_{\mathsf{w}} b(\mathsf{w}) \cdot \mathcal{I}(c,\mathsf{w}) = x$$
>
> but she is also certain that the chance function ch is such that:
>
> $$\sum_{\mathsf{w}} \mathrm{ch}(\mathsf{w}) \cdot \mathcal{I}(c,\mathsf{w}) = y, \ \text{ with } x \neq y$$
>
> in which case she violates the Principal Principle, then nonetheless:
>
> $$\mathsf{Est}_b(\mathcal{I}(c,\cdot)) = y$$
>
> That is, she ought to line up her own best estimate of c's inaccuracy with what she knows c's objective expected inaccuracy to be. (Konek and Levinstein, ms, p. 10, notation slightly altered).

So if she knows additional facts about how the world is set up, she must calculate the estimated inaccuracy using these.

Our 2b is given in a similar spirit to the principle Deference to Chance: Since she is introspective she has additional facts about how the world is set

---

[17] At least for the agenda $\{Heads, \neg Heads, \mathsf{P}^{\mathrm{Sophie}}\ulcorner Heads\urcorner, \mathsf{P}^{\mathrm{Sophie}}\ulcorner\neg Heads\urcorner\}$

up. In fact, she will know which world is actual. She should therefore only care about inaccuracy in that world. So we calculate her estimated inaccuracy only by considering that world, i.e.

$$\mathsf{Est}_b(\mathcal{I}(c,\cdot)) = \mathcal{I}(c,\mathsf{w}_b).$$

If we want to consider agents who aren't perfectly introspective or who are uncertain about facts in the world but still want to apply a criterion analogous to 2b, we will have to modify the definition of 2b to apply to non self-ref agendas.

7.3.2 How to measure inaccuracy in the general case − Deferring to the probabilistic modal structure

In Chapter 7 we assumed that agents that we considered were introspective and we restricted our attention to self-ref agendas, essentially assuming that the agent was omniscient in the relevant regards. In this chapter we will drop these assumptions. In Chapter 6 we already suggested that this is essentially assuming that the agent can be modelled by the trivial probabilistic modal structure $\mathfrak{M}_{\mathsf{omn}}$.
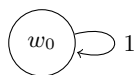
In this chapter we will still assume that the agent is modelled by *some* probabilistic modal structure but will not assume that it is $\mathfrak{M}_{\mathsf{omn}}$. We are not using these accuracy considerations to argue for the fact that the agent is appropriately modellable by such structures or which structure appropriately models an agent; that is something we are going to take as given. We can see this as a generalisation of the assumptions embedded in the previous considerations of these rationality criteria where we said the agent was introspective and omniscient so modellable by $\mathfrak{M}_{\mathsf{omn}}$. In future work we would like to consider dropping the assumption that we start off with a probabilistic modal structure.

We are therefore considering the question of:

> If an agent is appropriately represented as being at world w of a probabilistic modal structure $\mathfrak{M}$, what should her degrees of belief be like?

This is very connected to the question in Part I where we started with an $\mathfrak{M}$ modelling a situation and were trying to determine truth values of sentences. Here we start with some $\mathfrak{M}$ and try to determine the rationally required credences.

In Section 7.3 we proposed that if an agent is currently in credal state b, she should wish to be in the credal state which minimizes $\mathsf{Est}_b(\mathcal{I}(c,\cdot))$. For the case where we assume the agent is introspective and the agenda is a self-ref agenda, we proposed to define this as

$$\mathcal{I}(c,\mathsf{w}_b).$$

We will here provide a generalisation of this definition, where $\mathcal{I}(c,\mathsf{w}_b)$ is the special case for $\mathfrak{M}_{\mathsf{omn}}$.

In the general case we will ask the agent to defer to the meta-accessibility relation to calculate the estimated inaccuracy, using her current credences to interpret P at each world.

Self-ref agendas do not allow for any contingent sentences and we now want to drop this restriction on agendas. The other, fullness component of the self-ref agenda is still needed. For example if $\mathsf{P}\ulcorner\varphi\urcorner \geqslant r$ is in the agenda $\mathcal{A}$, then we also need that φ is. This is because in those cases it makes sense to just focus on such an agenda as it is self-contained in the appropriate way. We call agendas that have this feature full agendas.

Definition 7.3.3. We say that $\mathcal{A}$ is a *full agenda* if: for any $\mathrm{M} \in \mathsf{Mod}_{\mathcal{L}}^{\mathsf{PA,ROCF}}$ and $\mathrm{p}, \mathrm{p}' : \mathsf{Sent}_{\mathsf{P}} \to \mathbb{R}$,

$$\forall\varphi \in \mathcal{A}, (\mathrm{p}(\varphi) = \mathrm{p}'(\varphi)) \implies \forall\varphi \in \mathcal{A}, (\llbracket\varphi\rrbracket_{(\mathrm{M,p})} = \llbracket\varphi\rrbracket_{(\mathrm{M,p})})$$

What this says is that the truth values of all sentences in the agenda should depend on: firstly a choice of base model (this is the difference to the self-ref agendas) and secondly a choice of probability values for the sentences also in the agenda. This only rules out that it depends on the probability value assigned to a sentence not in the agenda.[18]

Since we want to allow the agent to have different degrees of belief at different worlds, to sum up how an alternative credal state looks from her current perspective we need to start off with her possible attitude *in each world*, as given by a prob-eval function, except restricted to an agenda.

Definition 7.3.4. An $\mathcal{A}$-*prob-eval function,* $\mathbf{c}$, is some collection of members of $\mathsf{Creds}_{\mathcal{A}}$, one for each $w \in W$. So $\mathbf{c}(w) : \mathcal{A} \to [0,1]$. We might also use the metavariable $\mathbf{b}$.

In the special case of self-ref agendas, we measured the inaccuracy of a credal function to w_b, because for self-ref agendas, truths were determined by a choice of credences. For full agendas we also need information about how to interpret the base language, so instead of measuring the inaccuracy to w_b, we measure it to $\mathsf{w}_{(\mathbf{M},\mathbf{b})(v)}$, which will be a member of $\mathsf{Worlds}_{\mathcal{A}}$. This will be given by the truth values when $\mathbf{M}(v)$ interprets the vocabulary from $\mathcal{L}$ and $\mathbf{b}(v)$ interprets P. These are then the true, ideal, or omniscient credences if the agent's degrees of belief, restricted to $\mathcal{A}$, are given by $\mathbf{b}$.

Definition 7.3.5. $\mathsf{w}_{(\mathbf{M},\mathbf{b})(v)}$ is a function from $\mathcal{A}$ to $\{0,1\}$ with

$$\mathsf{w}_{(\mathbf{M},\mathbf{b})(v)}(\varphi) := \llbracket\varphi\rrbracket_{\mathcal{M}_{(\mathbf{M}(v),\mathbf{b}(v))}} = \begin{cases} 1 & \mathcal{M}_{(\mathbf{M}(v),\mathbf{b}(v))} \models \varphi \\ 0 & \text{otherwise} \end{cases}$$

Where $\mathcal{M}_{(\mathbf{M}(v),\mathbf{b}(v))} \in \mathsf{Mod}_{\mathcal{L}_{\mathsf{P}}}^{\mathsf{PA,ROCF}}$ which, restricted to $\mathcal{L}$ is just $\mathbf{M}(v)$, and which has

$$\mathcal{M}_{(\mathbf{M}(v),\mathbf{b}(v))} \models \mathsf{P}\ulcorner\varphi\urcorner = r \iff \mathbf{b}(v)(\varphi) = r$$

For full agendas, the flexibility in the choice of $\mathcal{M}_{(\mathbf{M}(v),\mathbf{b}(v))}$ does not alter $\mathsf{w}_{(\mathbf{M},\mathbf{b})(v)}$.

We can now define the estimated inaccuracy from $\mathbf{b}$'s perspective.

---

[18]We would like to say: the agenda contains any sentences which it refers to the probability of, however $\mathsf{P}\ulcorner\varphi\urcorner \geqslant r \vee \mathsf{P}\ulcorner\varphi\urcorner < r$ itself forms full agenda because this sentence is true regardless of the probability value assigned to φ.

Definition 7.3.6. Let $\mathcal{A}$ be a full agenda and $\mathbf{b}$ an $\mathcal{A}$-prob-eval function. Define the inaccuracy of a credal state, $c \in \mathsf{Creds}_{\mathcal{A}}$, from $\mathbf{b}$-at-w's perspective, by:[19]

$$\mathsf{Est}_{\mathbf{b},w}(\mathcal{I}(c,\cdot)) := \sum_{v \in W} m_w\{v\} \cdot \mathcal{I}(c, \mathsf{w}_{(\mathbf{M},\mathbf{b})(v)}).$$

We can also measure the estimated inaccuracy of a whole $\mathcal{A}$-prob-eval function, $\mathbf{c}$, from $\mathbf{b}$'s perspective, by:

$$
\begin{aligned}
\mathsf{Est}_{\mathbf{b}}(\mathcal{I}(\mathbf{c},\cdot)) &:= \sum_{w \in W} \mathsf{Est}_{\mathbf{b},w}(\mathcal{I}(\mathbf{c}(w),\cdot)) \\
&= \sum_{w \in W} \sum_{v \in W} m_w\{v\} \cdot \mathcal{I}(\mathbf{c}(v), \mathsf{w}_{(\mathbf{M},\mathbf{b})(v)}).
\end{aligned}
$$

Of course, all this is assuming W is finite. If W is infinite, the analogous definitions can be made, they just become more complicated.

7.3.3 Connections to the revision theory

We have now suggested a proposal for how to measure inaccuracy. This was given by:

$$\mathsf{Est}_{\mathbf{b}}(\mathcal{I}(\mathbf{c},\cdot)) := \sum_{w \in W} \sum_{v \in W} m_w\{v\} \cdot \mathcal{I}(\mathbf{c}(v), \mathsf{w}_{(\mathbf{M},\mathbf{b})(v)})$$

Definition 7.3.7. $\mathcal{I}$ is an *additive and continuous strictly proper inaccuracy measure*, if there is some $\mathfrak{s} : [0,1] \times \{0,1\} \to [0,\infty]$ such that:

- $\mathfrak{s}(x,i)$ is a continuous function of x,

- $\mathfrak{s}$ is strictly proper, i.e. for all $r \in [0,1]$,

$$r \cdot \mathfrak{s}(x,1) + (1-r) \cdot \mathfrak{s}(x,0)$$

  is minimized uniquely (as a function of x) at r.

- $\mathcal{I}(c,\mathsf{w}) = \sum_{\varphi \in \mathcal{A}} \mathfrak{s}(c(\varphi), \mathsf{w}(\varphi))$.

This is an assumption that is usual when working with inaccuracy measures for the accuracy argument. There has been a body of work developing justifications for working with such rules (see, e.g., Joyce, 1998, 2009; Leitgeb and Pettigrew, 2010; Pettigrew, ms). Notably $\mathsf{AbsValDist}$ is *not* strictly proper, so we are ruling this out as a legitimate inaccuracy measure.

Proposition 7.3.8. *Suppose $\mathcal{A}$ is a (finite) full agenda, and $\mathcal{I}$ is an additive and continuous strictly proper inaccuracy measure. Then (due to the strict proprietry),*

$$\mathsf{Est}_{\mathbf{b},w}(\mathcal{I}(c,\cdot)) := \sum_{v \in W} m_w\{v\} \cdot \mathcal{I}(c, \mathsf{w}_{(\mathbf{M},\mathbf{b})(v)})$$

---

[19] $= \sum_{\mathsf{w} \in \mathsf{Worlds}_{\mathcal{A}}} m_w\{v \mid \mathsf{w}_{(\mathbf{M},\mathbf{b})(v)} = \mathsf{w}\} \cdot \mathcal{I}(c, \mathsf{w})$

is minimized uniquely at

$$c = \sum_{v \in W} m_w\{v\} \cdot \mathsf{w}_{(\mathbf{M},\mathbf{b})(v)},$$

i.e. for each $\varphi \in \mathcal{A}$,

$$c(\varphi) = m_w\{v \mid \mathsf{w}_{(\mathbf{M},\mathbf{b})(v)}(\varphi) = 1\}.$$

Therefore,

$$\mathsf{Est}_{\mathbf{b}}(\mathcal{I}(\mathbf{c}, \cdot))$$

is minimized uniquely at $\mathbf{c}$ with for each $w \in W$, $\varphi \in \mathcal{A}$,

$$\mathbf{c}(w)(\varphi) = m_w\{v \mid \mathsf{w}_{(\mathbf{M},\mathbf{b})(v)}(\varphi) = 1\}.$$

We have seen this characterisation before: we saw it in the definition of the revision sequence:

$$\mathbf{p}_{\alpha+1}(w)(\varphi) := m_w\{v \mid (\mathbf{M}, \mathbf{p}_\alpha)(v) \models \varphi\}$$

The difference is just in the use of the agenda. This is hopefully also what we would obtain by applying the inaccuracy considerations to the (infinite) agenda Sent_P. The complication the agenda brings in should be studied, but we will not do that here.

So in fact, the successor definition of probability in the revision sequence can be seen to be the appropriate way to minimize inaccuracy: suppose the agent is currently in some epistemic state. To minimize her estimated inaccuracy she would like to be in a (generally) different state. The step from the first state to the second is a *revision* of her credal state. Once she has moved her credences she will again like to move, which is another revision step. The revision theory of probability characterises these steps in minimizing estimated accuracy.

This is very different to the traditional setup because if we have made these assumptions then there is no credal state which minimizes its own estimated inaccuracy, or *looks best from its own perspective*. I.e. there are no *stable states*. However, to obtain stable states, or fixed points, one can instead consider non-classical evaluation schemes and probability functions which do not always assign single real numbers to each proposition.

7.3.4 Non-classical accuracy criteria and connections to Kripkean semantics

In Chapters 3 and 4 we considered semantics that were based on non-classical evaluations schemes so that by revising one's probabilities one obtains a fixed point. This resulted in a final semantics where one can interpret the probabilities as being assigned by a range instead of a single point, or by a set of probability values. The other way in which the Kripkean semantics requires alterations to the rationality requirements is their reliance on non-classical logic. We therefore have to alter the measuring-inaccuracy framework to account for these alterations.

Connections to Chapter 3

There has been work on applying the rational constraints of Dutch book and accuracy considerations to non-classical semantics (Williams, 2014, 2012b,a; Paris, 2001). In the accuracy considerations we had that $\mathsf{w} : \mathcal{A} \to \{0,1\}$ where $\mathsf{w}(\varphi)$ is the numerical representation of the truth value of φ in w. Williams (2012b) shows that also in the general, non-classical setting, accuracy considerations require one to pick some credal state in the convex hull of the ws, even when these assign numerical representations of the truth values according to some non-classical evaluation schema and also when the truth value representations are not in $\{0,1\}$ but are some other real numbers. These *do* require one to assign some particular number as the numerical representation of the truth value, and from these one obtains credal states that assign a single number as a degree of belief to each sentence. These credal states may not be (classically) probabilistic, but they are instead non-classical probabilities. We will slightly generalise this framework to obtain something like the ranges we considered in Chapter 3.

We will first mention how the agenda can play a role working with evaluation functions instead of prob-eval functions, but the role is similar.

Proposition 7.3.9. *If $\mathcal{A}$ is a full agenda, then for any $\mathfrak{M}$[20] and evaluation functions f, f',[21]*

$$\forall \varphi \in \mathcal{A}, \ \forall w \in W, \quad \#\varphi \in f(w) \iff \#\varphi \in f'(w)$$
$$\implies \quad \forall \varphi \in \mathcal{A}, \ \forall w \in W, \quad (w, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \varphi \iff (w, f') \models^{\mathsf{SKP}}_{\mathfrak{M}} \varphi$$

So for a full agenda, just having the information about the evaluation of sentences in the agenda is enough to characterise Θ.

Definition 7.3.10. An $\mathcal{A}$-*evaluation function*, g, assigns to each $w \in W$ some subset of $\mathcal{A}$.

We will use Θ as a function from $\mathcal{A}$-evaluation functions to $\mathcal{A}$-evaluation functions using the obvious modification of the definition from Definition 3.2.5.[22]

In the Kripkean semantics based on three-valued Strong Kleene evaluation scheme, so where we assumed f was consistent, we considered credal states as assigning intervals of real numbers to each sentences. These were given by

$$\underline{\mathrm{p}_{(v,f)}}(\varphi) := \sup\{\alpha \mid \mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{P}_{\geqslant}(\ulcorner\varphi\urcorner, \ulcorner\alpha\urcorner)\}$$
$$= m_w\{v \mid (v, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \varphi\}$$
$$\overline{\mathrm{p}_{(v,f)}} := \inf\{\alpha \mid \mathsf{IM}_{\mathfrak{M}}[w, f] \models \mathsf{P}_{<}(\ulcorner\varphi\urcorner, \ulcorner\alpha\urcorner)\}$$
$$= m_w\{v \mid (v, f) \not\models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\varphi\}$$

These values can in fact be justified by accuracy considerations. For this framework to work it is important here that we fix *three-valued* strong Kleene evaluation scheme and assume that f is consistent.[23] We are then working with

---

[20]With each $\mathbf{M}(w) \in \mathsf{Mod}^{\mathsf{PA},\mathsf{ROCF}}_{\mathcal{L}}$.

[21]Which implies, $\forall \varphi \in \mathcal{A}, \ \forall w \in W, \ \#\varphi \in \Theta(f) \iff \#\varphi \in \Theta(f')$

[22]I.e. Let $\Theta(g)(w) := \{\varphi \in \mathcal{A} \mid \#\varphi \in \Theta(f)(w)\}$ for some evaluation function f, where $\#\varphi \in f(w) \iff \varphi \in g(w)$.

[23] An analysis of how this restriction could be dropped would be very interesting but is left to future research. See Janda (2016) for some results on applying accuracy arguments in Belnap's 4-valued logic setting, which is the setting we'd be in by considering the Kripke construction allowing for non-consistent evaluation functions.

three different truth values and we will consider two different weightings: a lowerweighting and an upperweighting.

truth value	characterised by	given weight	
		lowerweighting	upperweighting
φ true	$(v, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \varphi$	1	1
φ neither	$(v, f) \not\models^{\mathsf{SKP}}_{\mathfrak{M}} \varphi$ and $(v, f) \not\models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\varphi$	0	1
φ false	$(v, f) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\varphi$	0	0

These two different weightings characterise the extremities of natural weightings that can be assigned to the truth value neither.

Definition 7.3.11. Fix a full agenda $\mathcal{A}$. For $\varphi \in \mathcal{A}$, and an $\mathcal{A}$-evaluation function g, let:

$$\underline{\mathsf{w}_{(v,g)}}(\varphi) := \begin{cases} 1 & (v, g) \models^{\mathsf{SKP}}_{\mathfrak{M}} \varphi \\ 0 & (v, g) \not\models^{\mathsf{SKP}}_{\mathfrak{M}} \varphi \text{ and } (v, g) \not\models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\varphi \\ 0 & (v, g) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\varphi \end{cases}$$

$$\overline{\mathsf{w}_{(v,g)}}(\varphi) := \begin{cases} 1 & (v, g) \models^{\mathsf{SKP}}_{\mathfrak{M}} \varphi \\ 1 & (v, g) \not\models^{\mathsf{SKP}}_{\mathfrak{M}} \varphi \text{ and } (v, g) \not\models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\varphi \\ 0 & (v, g) \models^{\mathsf{SKP}}_{\mathfrak{M}} \neg\varphi \end{cases}$$

These are the truth value assignments corresponding to the upper and lower weightings.

Proposition 7.3.12. *Let $\mathcal{I}$ be some additive and continuous strictly proper inaccuracy measure. Let $\mathcal{A}$ be a full agenda, and g an $\mathcal{A}$-evaluation function, then:*

- $\underline{\mathrm{p}_{(w,\Theta(g))}}$ *uniquely minimizes*

$$\sum_{v \in W} m_w\{v\} \mathcal{I}(c, \underline{\mathsf{w}_{(v,g)}})$$

  with respect to c.

- $\overline{\mathrm{p}_{(w,\Theta(g))}}$ *uniquely minimizes*

$$\sum_{v \in W} m_w\{v\} \mathcal{I}(c, \overline{\mathsf{w}_{(v,g)}})$$

  with respect to c.

And as in the discussion of Section 7.3.3, this may be seen as the appropriate way to measure the estimated inaccuracy of a credal state.

Proposition 7.3.13. *Suppose $\mathcal{A}$ is a full agenda taking the form $\{\varphi_1, \neg\varphi_1, \ldots, \varphi_n, \neg\varphi_n\}$. If $\mathcal{I}$ is normal and g_1 and g_2 are consistent, then:*

$$\mathcal{I}(\underline{\mathrm{p}_{(v,g_2)}}, \underline{\mathsf{w}_{(v,g_1)}}) = \mathcal{I}(\overline{\mathrm{p}_{(v,g_2)}}, \overline{\mathsf{w}_{(v,g_1)}})$$

191

Proof. If g_1 and g_2 are consistent,

$$\overline{\mathrm{p}_{(v,g_2)}}(\varphi) = 1 - \underline{\mathrm{p}_{(v,g_2)}}(\neg\varphi)$$

$$\overline{\mathsf{w}_{(v,g_1)}}(\varphi) = 1 - \underline{\mathsf{w}_{(v,g_1)}}(\neg\varphi)$$

$$\overline{\mathrm{p}_{(v,g_2)}}(\neg\varphi) = 1 - \underline{\mathrm{p}_{(v,g_2)}}(\varphi)$$

$$\overline{\mathsf{w}_{(v,g_1)}}(\neg\varphi) = 1 - \underline{\mathsf{w}_{(v,g_1)}}(\varphi).$$

So we have equalities between the relevant multisets.

$$\{|\overline{\mathrm{p}_{(v,g_2)}}(\varphi_i) - \overline{\mathsf{w}_{(v,g_1)}}(\varphi_i)| \,|\, i = 1, \ldots, n\}$$
$$\cup \{|\overline{\mathrm{p}_{(v,g_2)}}(\neg\varphi_i) - \overline{\mathsf{w}_{(v,g_1)}}(\neg\varphi_i)| \,|\, i = 1, \ldots, n\}$$
$$= \{|\underline{\mathrm{p}_{(v,g_2)}}(\neg\varphi_i) - \underline{\mathsf{w}_{(v,g_1)}}(\neg\varphi_i)| \,|\, i = 1, \ldots, n\}$$
$$\cup \{|\underline{\mathrm{p}_{(v,g_2)}}(\varphi_i) - \underline{\mathsf{w}_{(v,g_1)}}(\varphi_i)| \,|\, i = 1, \ldots, n\}$$

So if $\mathcal{I}$ is normal we have our result. $\qquad\square$

Definition 7.3.14. For $\mathcal{A}$ a full agenda, $\mathcal{I}$ satisfying normality and g, g' consistent $\mathcal{A}$-evaluation functions, define:

$$\mathsf{Est}_g(\mathcal{I}(g', \cdot)) := \sum_{w \in W} \sum_{v \in W} m_w\{v\} \cdot \mathcal{I}(\underline{\mathrm{p}_{(v,g)}}, \underline{\mathsf{w}_{(v,g')}})$$

Proposition 7.3.15. *For $\mathcal{A}$ a finite, full agenda, g a consistent $\mathcal{A}$-evaluation function and $\mathcal{I}$ an inaccuracy measure satisfying StrictProprietry and Normality, we have:*

$$\Theta(g) = \underset{g' \ consistent}{\arg\min} \left(\mathsf{Est}_g(\mathcal{I}(g', \cdot)) \right)$$

It is common in the traditional accuracy criterion to only be deemed irrational if the dominating credal state is immodest.[24] In this case, the dominating credal state may not be immodest (since it would advise moving to $\Theta(\Theta(f))$), however after a sequence of such moves one will reach a fixed point, which is a credal state that is immodest and looks best from its own perspective.

Proposition 7.3.16. *Suppose $\mathcal{A}$ a finite, full agenda, g is a consistent $\mathcal{A}$-evaluation function and $\mathcal{I}$ an inaccuracy measure satisfying StrictProprietry and Normality.*

g uniquely minimizes $\mathsf{Est}_g(\mathcal{I}(g', \cdot))$ with respect to g' if and only if g is some fixed point evaluation function restricted to $\mathcal{A}$.

This is different from the revision sequence where we ended up with a never ending sequence of moves trying to minimise estimated inaccuracy.

Idea for a connection to Chapter 4

In Chapter 4 we considered a variant of probability where we can assign a set of probabilities to be the interpretation of P. We therefore interpret P using *imprecise probabilities*. There has recently been work in trying to apply accuracy considerations to imprecise probabilities, and particularly to try to

---

[24]See Pettigrew (ms) for a discussion.

find an appropriate inaccuracy measure for them (see Seidenfeld et al., 2012; Mayo-Wilson and Wheeler, ta; Schoenfield, 2015). It has turned out that one cannot find an inaccuracy measure that assigns single real numbers and which satisfies StrictProprietry.

Here, we suggest an alternative way of measuring the inaccuracy where we let the inaccuracy be a set of numbers, namely the set of inaccuracy of the members of the credal set. This works point-wise and collects the results. Suppose $\mathcal{C}$ is a collection of members of $\mathsf{Creds}_\mathcal{A}$. Define:

$$\mathcal{I}(\mathcal{C}, \mathsf{w}) := \{\mathcal{I}(c, \mathsf{w}) \mid c \in \mathcal{C}\}$$

To try and apply this to our framework we have not just imprecise probabilities, but imprec-prob-eval functions which also account for the different worlds in the probabilistic modal structure. We define $\mathcal{A}$-imprec-prob-eval functions in an analogous way to $\mathcal{A}$-prob-eval functions in Definition 7.3.4.

We might hope to define estimated inaccuracy in a similar manner, by:

$$\mathsf{Est}_{\mathcal{B}}(\mathcal{I}(\mathcal{C}, \cdot)) := \{\mathsf{Est}_{\mathbf{b}}(\mathcal{I}(\mathbf{c}, \cdot))) \mid \mathbf{b} \in \mathcal{B}, \mathbf{c} \in \mathcal{C}\}$$

We would then hope that the way to minimize this matches the semantics. It is not immediately clear what it means to minimize a set of numbers. But we can see the following connection.

Suppose $\mathcal{A}$ is a finite full agenda. Then Θ as defined in Chapter 4 can also be conceived of as a function from $\mathcal{A}$-imprec-prob-eval functions to $\mathcal{A}$-imprec-prob-eval functions. Suppose $\mathcal{I}$ is an inaccuracy measure satisfying StrictProprietry. Then for $\mathcal{B}$ an $\mathcal{A}$-imprec-prob-eval function we have:

$$\Theta(\mathcal{B}) = \left\{ \arg\min_{\mathbf{c}}(\mathsf{Est}_{\mathbf{b}}(\mathcal{I}(\mathbf{c}, \cdot))) \,\middle|\, \mathbf{b} \in \mathcal{B} \right\}$$

So this should somehow be able to be construed as the result of minimizing estimated inaccuracy, where inaccuracy is measured by a set of real numbers. The details of exactly how this works are yet to be worked out. We would then have that the *stable* states are exactly those that minimize estimated inaccuracy from their own perspective.

We can now see the feature that we mentioned in Chapter 4: At a fixed point (where $\mathcal{B} = \Theta(\mathcal{B})$), for every $\mathbf{b} \in \mathcal{B}$ there is some $\mathbf{c} \in \mathcal{B}$ such that $\mathbf{c}$ looks best from $\mathbf{b}$'s perspective, i.e. it minimizes $\mathsf{Est}_{\mathbf{b}}(\mathcal{I}(\mathbf{c}', \cdot))$ with respect to $\mathbf{c}'$.

In conclusion, by modifying the measuring inaccuracy framework in different ways we obtain the result that the different semantics that we have developed can each be seen as closely relating to minimising estimated inaccuracy. Though certain details still need to be worked out, this is a very interesting connection.

Appendix 7.A Minimize Self-Inaccuracy's flexibility to get definable regions without Normality

In this thesis we have been working in formal languages which generally aren't able to express all regions, so the result in Theorem 7.1.11 doesn't apply to the languages we've been considering. In the following result we make further

assumptions on SelfInacc in order to derive a result that will apply in these restricted languages.

Theorem 7.A.1. *Let $n \geqslant 2$. $\mathbf{x}$ is here a metavariable for some member of $\mathbb{R}^n$. Suppose:*

- *$\mathcal{I}$ satisfies Extensionality,*

- *$\mathcal{I}$ is a continuous function wrt the Euclidean topology, e on $\mathbb{R}$,*

- *$\mathcal{I}$ is bounded: For each $\mathsf{w} \in \mathsf{Worlds}_{\mathcal{A}}$, $\mathcal{I}(c, \mathsf{w})$ does not go to 0 as $c(\varphi)$ goes to infinity for some $\varphi \in \mathcal{A}$. I.e.:*

    - *There is some $q_0 > 0$ and $q_1 > 0$ such that for all $\mathsf{w} \in \mathsf{Worlds}_{\mathcal{A}}$, $\mathsf{wMoreAcc}_{q_1}(\mathsf{w}) \subseteq [-q_0, q_0]^n$.*

- *There is no $\mathbf{x}$ with both $\mathcal{I}(\mathbf{x}, \mathsf{w}_n^{\mathsf{pos}}) = 0$ and $\mathcal{I}(\mathbf{x}, \mathsf{w}_n^{\mathsf{neg}}) = 0$.*

Then there is some $r > 0$ such that for all $\mathbf{x}$ with $\mathcal{I}(\mathbf{x}, \mathsf{w}_n^{\mathsf{pos}}) \leqslant r$ or $\mathcal{I}(\mathbf{x}, \mathsf{w}_n^{\mathsf{neg}}) \leqslant r$, there is a region $R \subseteq \mathbb{R}^n$ and a sentence δ of $\mathcal{L}_{\mathsf{P}_{\geqslant r}}$ with

$$\delta \leftrightarrow \text{``}c \in R\text{''}$$

such that SelfInacc is minimised uniquely at $\mathbf{x}$.

If, in addition, $\mathcal{I}$ satisfies:

- *$\mathcal{I}(\mathsf{w}_n^{\mathsf{pos}}, \mathsf{w}_n^{\mathsf{pos}}) = \mathcal{I}(\mathsf{w}_n^{\mathsf{neg}}, \mathsf{w}_n^{\mathsf{neg}}) = 0$*

Then there is some $r' > 0$ such that for all $\mathbf{x}$ with $e(\mathbf{x}, \mathsf{w}_n^{\mathsf{pos}}) \leqslant r'$ or $e(\mathbf{x}, \mathsf{w}_n^{\mathsf{neg}}) \leqslant r'$, there is a region $R \subseteq \mathbb{R}^n$ and a sentence δ of $\mathcal{L}_{\mathsf{P}_{\geqslant r}}$ with

$$\delta \leftrightarrow \text{``}c \in R\text{''}$$

such that SelfInacc(c) is minimised uniquely at $\mathbf{x}$.

Proof. Fix some $n \geqslant 2$ We first define the rectangular approximation of the ball around $\mathsf{w}_n^{\mathsf{neg}}$. For $i_1, \ldots i_n \in \mathbb{Z}$, $m \in \mathbb{N}$ let $I_m^{i_1, \cdots i_n} = ([i_1/2^m, i_1+1/2^m] \times \ldots \times [i_n/2^m, i_n+1/2^m])$. Let

$$\mathsf{RectClB}_r^m(w) := \bigcup_{\substack{i_1, \ldots, i_n \text{ such that} \\ \mathsf{wMoreAcc}_r(w) \cap I_m^{i_1, \ldots, i_n} \neq \varnothing}} I_m^{i_1, \ldots, i_n}$$

Observe that

$$\bigcap_{m \in \mathbb{N}} \mathsf{RectClB}_r^m(w) = \mathsf{wMoreAcc}_r(w)$$

since $\mathcal{I}$ is continuous.[25]

For $m' \geqslant m$ and $r' \geqslant r$,

$$\mathsf{RectClB}_r^m(\mathsf{w}_n^{\mathsf{neg}}) \supseteq \mathsf{RectClB}_{r'}^{m'}(\mathsf{w}_n^{\mathsf{neg}})$$

---

[25]Suppose $\mathbf{x} \in \bigcap_{m \in \mathbb{N}} \mathsf{RectClB}_r^m(w)$. Then for each $m \in \mathbb{N}$ there is some $I_m^{i,j} \ni \mathbf{x}$ with some $\mathbf{x}_m \in \mathsf{wMoreAcc}_r(w)$ and $\mathbf{x}_m \in I_m^{i,j}$. So $e(\mathbf{x}, \mathbf{x}_m) \leqslant \sqrt{\frac{1}{m^n}}$. Since $\mathsf{wMoreAcc}_r(w)$ is closed it must therefore be that $\mathbf{x} \in \mathsf{wMoreAcc}_r(w)$.

and
$$\mathsf{wMoreAcc}_r(\mathsf{w}_n^{\mathsf{pos}}) \supseteq \mathsf{wMoreAcc}_{r'}(\mathsf{w}_n^{\mathsf{pos}}).$$

By the assumption that $\mathcal{I}$ is bounded we have that there are $q_0, q_1 > 0$ with $\mathsf{wMoreAcc}_{q_1}(\mathsf{w}_n^{\mathsf{neg}}) \cap \mathsf{wMoreAcc}_{q_1}(\mathsf{w}_n^{\mathsf{pos}}) \subseteq [-q_0, q_0]^n$, and therefore for all $r \leqslant q_1$,

$$\mathsf{RectCIB}_r^m(\mathsf{w}_n^{\mathsf{neg}}) \cap \mathsf{wMoreAcc}_r(\mathsf{w}_n^{\mathsf{pos}}) \subseteq [-q_0 - 2, q_0 + 2]^n$$

We also have:

$$\bigcap_{\substack{r \leqslant q_1 \\ m \in \mathbb{N}, m > 0}} \mathsf{RectCIB}_r^m(\mathsf{w}_n^{\mathsf{neg}}) \cap \mathsf{wMoreAcc}_r(\mathsf{w}_n^{\mathsf{pos}}) = \mathsf{wMoreAcc}_0(\mathsf{w}_n^{\mathsf{neg}}) \cap \mathsf{wMoreAcc}_0(\mathsf{w}_n^{\mathsf{pos}}) = \varnothing.$$

And since $\mathcal{I}$ is continuous, each of these sets is closed.

$[-q_0 - 2, q_0 + 2]^n$ is compact, so by the finite intersection property of compact sets, there is some $q \geqslant q_1, M \in \mathbb{N}$ such that

$$\mathsf{RectCIB}_q^M(\mathsf{w}_n^{\mathsf{neg}}) \cap \mathsf{wMoreAcc}_q(\mathsf{w}_n^{\mathsf{pos}}) = \varnothing$$

Now take any $\mathbf{x}$ with $\mathcal{I}(\mathbf{x}, \mathsf{w}_n^{\mathsf{pos}}) < q$. So $\mathbf{x} \in \mathsf{wMoreAcc}_q(\mathsf{w}_n^{\mathsf{pos}})$.
Take

$$R := \mathsf{RectCIB}_q^M(\mathsf{w}_n^{\mathsf{neg}}) \cup \{\mathbf{x}\}.$$

Define a sentence δ with a δ-agenda $\delta \leftrightarrow$ "$c \in R$". This can be done by using the diagonal lemma with the formula:

$$\bigvee_{\substack{\mathsf{wMoreAcc}_r(w) \cap I_m^{i_1,\dots,i_n} \neq \varnothing \\ |i_j \cdot m| < q_0 \text{ for } j=1,\dots,n}} \left(\bigwedge_{j=1,\dots,n} (i_j/m \leqslant \mathsf{P}(\overbrace{v \veebar (v \veebar (\dots (v \veebar (v \veebar v)\dots)))}^{j}) \leqslant i_j+1/m) \right)$$

$$\vee \bigwedge_{k=1,\dots,n} \mathsf{P}(\overbrace{v \veebar (v \veebar (\dots (v \veebar (v \veebar v)\dots)))}^{j}) = x_k$$

Observe $\mathsf{SelfInacc}(\mathbf{x}) \leqslant q$. So

$$\mathsf{wMoreAcc}_{\mathsf{SelfInacc}(\mathbf{x})}(w) \subseteq \mathsf{wMoreAcc}_q(w) \subseteq \mathsf{RectCIB}_q^M(w)$$

for $w \in \{\mathsf{w}_n^{\mathsf{pos}}, \mathsf{w}_n^{\mathsf{neg}}\}$. Therefore

$R \cap \mathsf{wMoreAcc}_{\mathsf{SelfInacc}(\mathbf{x})}(\mathsf{w}_n^{\mathsf{pos}})$
$\subseteq (\{\mathbf{x}\} \cap \mathsf{wMoreAcc}_{\mathsf{SelfInacc}(q)}(\mathsf{w}_n^{\mathsf{pos}})) \cup (\mathsf{RectCIB}_q^M(\mathsf{w}_n^{\mathsf{neg}}) \cap \mathsf{wMoreAcc}_{\mathsf{SelfInacc}(q)}(\mathsf{w}_n^{\mathsf{pos}}))$
$\subseteq \{\mathbf{x}\}$

and
$$R \supseteq \mathsf{RectCIB}_q^M(\mathsf{w}_n^{\mathsf{neg}}) \supseteq \mathsf{wMoreAcc}_{\mathbf{x}}(\mathsf{w}_n^{\mathsf{neg}})$$

So by Proposition 7.1.5, $\mathbf{x}$ minimizes $\mathsf{SelfInacc}$.

Now take any $\mathbf{x}$ with $\mathcal{I}(\mathbf{x}, \mathsf{w}_n^{\mathsf{neg}}) < q$. Then we can take $R := \mathsf{RectCIB}_q^M(\mathsf{w}_n^{\mathsf{neg}}) \setminus \{\mathbf{x}\}$ and apply analogous reasoning. $\qquad\square$

One might have hoped that such accuracy considerations could lead to support for some generalized probabilism, in the sense of Williams (2014), or some other nice constraints on the rationally required credal states. But I think this shows that this is not possible since for any $\mathbf{x}$ sufficiently $\mathcal{I}$-close to $\mathsf{w}_n^{\mathsf{pos}}$ we can pick a sentence where $\mathbf{x}$ will minimize $\mathsf{SelfInacc}(c)$.

Corollary 7.A.2. *If $\mathcal{I}$ satisfies the above then we can always find some δ where* $\mathsf{SelfInacc}(c)$ *is minimized by some non-probabilistic* $\mathbf{x}$.

Proof. Using the above theorem we can find an N with the required properties. Since $\mathcal{I}$ is truth directed we can therefore pick some $\mathbf{x}$ non-probabilistic with $\mathcal{I}(\mathbf{x}, \mathsf{w}_n^{\mathsf{pos}}) < \frac{1}{N}$ which suffices. $\qquad\square$

We should also be able to drop the condition that $\mathcal{I}$ is truth directed to some extent, though continuity by itself is not enough. Continuity is essential to find the splitting regions.

Chapter 8

Dutch book Criterion

8.1 Introduction

We shall now consider the Dutch book argument. This says that a rational agent should have a credal state which has the feature that if he buys bets at prices governed by his credences then he will not buy a bet that will guarantee him a monetary loss.[1]

In the framework that we are working in, there are sentences that talk about probabilities. This means that the truth of such sentences depends on what the agent's credences are. Moreover we have sentences that talk about their *own* probabilities, such as π, whose truth therefore depends on what the agent's credences are in that very sentence. What this can then lead to is the result that whatever the agent's credence in π is, if he bets in accordance with that credence then there is a bet he will accept but which will lead him to guaranteed loss of money.

So even if an agent values money positively and linearly and values nothing else, he might be better not to bet at his credences. There are other cases where it has been argued that an agent's betting odds are not, or should not be, identical to his degrees of belief even when the agent values money positively and linearly and values nothing else. For example if the agent has less information than the bookie (Talbott, 1991), the agent's utilities are not linear in money (Seidenfeld et al., 1990; Maher, 1993), he will be irrational at a future time (Christensen, 1991), he is aware that the size or existence of a bet is correlated with the truth of the propositions the bet is in (Bradley and Leitgeb, 2006),[2] or when the proposition is unverifiable (Weatherson, 2003). We will add the additional case: when the truth value of the sentence the agent is betting on depends on his credences. It is a particularly non-pragmatic case of when betting odds and degrees of belief may fall apart.

What we will do in this chapter is to see how far we can go under the assumption that the agent *does* bet with his credences and try to determine what the resulting rationally required credences are.[3] In developing this criterion we

---

[1] Hájek (2008) can be consulted for an introduction to Dutch book arguments.

[2] These different cases are presented in Bradley and Leitgeb (2006).

[3] This could also be seen as an analysis of what the agent's fair betting odds should be if there are sentences which talk about his betting odds. So instead of considering a language which can talk about the agent's credences we instead consider a language which can talk

will be expanding on ideas from Caie (2013). Caie does not think that Dutch book arguments in fact have any normative force and I will not be arguing that they do, but it is nonetheless an interesting question to consider for *if* one thinks that they do have normative force. Certain interesting features may also arise from considering this, for example, there turns out to be a connection between self-inaccuracy and minimizing possible overall Dutch book losses (Section 8.5). Since we want to reject both, it would in the future be interesting to see if how the justification for rejecting them can be connected.

In this chapter we first present this result that any credal state can be Dutch booked (Section 8.2). In Section 8.3 we consider how one can find a Dutch book criterion that doesn't fall silent too often and in Section 8.4 we present our suggested proposal. This is that an agent should minimize her overall guaranteed losses, which we understand by her minimizing her *average* guaranteed losses on bets where there is just £1 at stake. In Section 8.5 we present the aforementioned result that this criterion turns out to be the same as Minimize Self-Inaccuracy using the inaccuracy measure AbsValDist. Finally in Section 8.6 we return to the possibility that this is a case where an agent should not bet with his credences, discussing how the agent's credences should determine her betting odds.

8.2   Any credal state can be Dutch booked

We write a bet on φ with stake r as (r, φ). This is the bet that gives the agent £r if φ is true, and nothing if φ is false. For example for the bet $(-1, \varphi)$, if φ is true then the agent will receive £-1, i.e. have to pay £1.

The possible sets of bets over the agenda $\mathcal{A} = \{\varphi_1, \ldots, \varphi_n\}$ are of the form $\mathcal{B} = \{(r_1, \varphi_1), \ldots, (r_n, \varphi_n)\}$ for any $r_i \in \mathbb{R}$. A Dutch book is some set of bets where if the agent considers each individually he will decide to pay a certain amount for each, but where if he pays that much the bets together lead him to be guaranteed-ly out-of-pocket.

To formulate this, we will assume that the agent evaluates the individual bets by using his credences, so an agent who has $c(\varphi) = 0.7$ will be willing to pay any amount of money less than £0.70 for the bet $(1, \varphi)$. More generally: an agent who has credences c and bets with them will be willing to pay any amount less than £$rc(\varphi)$ to receive the bet (r, φ). He is indifferent between paying £$rc(\varphi)$ to receive the bet (r, φ) or not paying anything and not receiving anything. We will make the simplifying assumption that he will always take a bet at the fair-prices, which are the prices at which he is indifferent between buying the bet or not. This assumption won't play an essential role in the discussion but will allow us to set things up more easily.

To keep the language simple we assume that he will also take collections of such bets at the same prices. He will therefore pay anything less than £$\sum_{i=1,\ldots,n} r_i c(\varphi_i)$ to receive the collection of bets $\{(r_1, \varphi_1), \ldots, (r_n, \varphi_n)\}$. However, the way we think this should be understood is not that he will actually evaluate the collection at that price but instead that this measures the collection of the individually evaluated prices, and a Dutch book shows the incoherence

---

about the agent's fair betting odds. The question then becomes: consider the sentence, $\pi_{\mathsf{Betting}} \leftrightarrow \mathsf{FairBettingOdds}\ulcorner \pi_{\mathsf{Betting}} \urcorner < 1/2$. What should the agent's fair betting price in π_{Betting} be?

of an agent's judgements of value: he evaluates the collection of these bets to be worth less than the sum of the values of the individual bets.

If our agent has bought the collection of bets $\{(r_1, \varphi_1), \ldots, (r_n, \varphi_n)\}$, and w is the actual world, then he will win £$\sum_{i=1,\ldots,n} r_i \mathsf{w}(\varphi_i)$. The limit of his potential loss on this collection of bets in w are therefore given by

$$\text{Amount paid} - \text{Amount won}$$
$$= \sum_{i=1,\ldots,n} r_i c(\varphi_i) - \sum_{i=1,\ldots,n} r_i \mathsf{w}(\varphi_i)$$
$$= \sum_{i=1,\ldots,n} r_i(c(\varphi_i) - \mathsf{w}(\varphi_i))$$

We therefore define:

Definition 8.2.1. Let $\mathcal{A} = \{\varphi_1, \ldots \varphi_n\}$. For $c \in \mathsf{Creds}_{\mathcal{A}}$, $\mathsf{w} \in \mathsf{Worlds}_{\mathcal{A}}$, and $r_1, \ldots, r_n \in \mathbb{R}$, define:

$$\mathsf{Loss}_{\{(r_1, \varphi_1), \ldots, (r_n, \varphi_n)\}}(c, \mathsf{w}) = \sum_{i=1,\ldots,n} r_i(c(\varphi_i) - \mathsf{w}(\varphi_i))$$

Here we are using $\mathsf{Worlds}_{\mathcal{A}}$ as defined in Definition 6.2.6, i.e.

$$\mathsf{Worlds}_{\mathcal{A}} = \{\mathcal{M} {\restriction}_{\mathcal{A}} \mid \mathcal{M} \in \mathsf{Mod}_{\mathcal{L}_{\mathsf{P}}}^{\mathsf{PA},\mathsf{ROCF}}\}.$$

Definition 8.2.2. Suppose $\mathcal{A} \subseteq \mathsf{Sent}_{\mathcal{L}}$ and $c \in \mathsf{Creds}_{\mathcal{A}}$. A *Dutch book* against c is some $\mathcal{B}$, such that

$$\text{for each } \mathsf{w} \in \mathsf{Worlds}_{\mathcal{A}}, \ \mathsf{Loss}_{\mathcal{B}}(c, \mathsf{w}) > 0.$$

I.e. it is given by some $\varphi_1, \ldots, \varphi_n \in \mathcal{A}$ and $r_1, \ldots, r_n \in \mathbb{R}$ such that

$$\text{for each } \mathsf{w} \in \mathsf{Worlds}_{\mathcal{A}}, \ \mathsf{Loss}_{\{(r_1, \varphi_1), \ldots, (r_n, \varphi_n)\}}(c, \mathsf{w}) > 0$$

Here we have only defined Dutch books for agendas that do not depend on the agent's credences.[4] In Definition 8.2.4 we will define Dutch books for self-ref agendas.[5]

The standard Dutch book rational requirement says that an agent should not have credences against which there is a Dutch book. We state that here, again just for agendas that do not depend on the agent's credences.

Rationality Criterion 3 (Usual Dutch Book Criterion). *Suppose $\mathcal{A} \subseteq \mathsf{Sent}_{\mathcal{L}}$ and $c \in \mathsf{Creds}_{\mathcal{A}}$. An agent should not have credences c if there is a Dutch book against c.*

I.e. Fix $\mathcal{A} = \{\varphi_1, \ldots, \varphi_n\} \subseteq \mathsf{Sent}_{\mathcal{L}}$. If there are some $r_1, \ldots, r_n \in \mathbb{R}$ such that for all $\mathsf{w} \in \mathsf{Worlds}_{\mathcal{A}}$,[6]

$$\mathsf{Loss}_{\{(r_1, \varphi_1), \ldots, (r_n, \varphi_n)\}}(c, \mathsf{w}) > 0$$

then $c \in \mathsf{Creds}_{\mathcal{A}}$ is irrational.

---

[4]We formalised this by restricting it to agendas which are $\subseteq \mathsf{Sent}_{\mathcal{L}}$. In fact we could apply the same criterion whenever $\mathcal{A}$ *doesn't depend on* P, i.e. if $\mathcal{M}, \mathcal{M}' \in \mathsf{Mod}_{\mathcal{L}_{\mathsf{P}}}^{\mathsf{PA},\mathsf{ROCF}}$ such that for all $\varphi \in \mathsf{Sent}_{\mathcal{L}}$, $\mathcal{M} \models \varphi \iff \mathcal{M}' \models \varphi$, then for all $\varphi \in \mathcal{A}$, $\mathcal{M} \models \varphi \iff \mathcal{M}' \models \varphi$. This would, for example, also allow $\mathsf{P}\ulcorner\varphi\urcorner \geqslant r \vee \mathsf{P}\ulcorner\varphi\urcorner < r$ to be a sentence in $\mathcal{A}$. We do not do this because it would just add complication without much additional benefit.

[5]This does not cover all agendas and a general characterisation of a Dutch book is desirable, but we do not yet have that.

[6]We could instead define this to be $> \epsilon$ for some $\epsilon > 0$, however this won't essentially effect our discussion here.

There is a well-known that says that there is a Dutch book against c if and only if c fails to satisfy the axioms of probability (over PA and ROCF, since we have only allowed models of these theories to give $\text{Worlds}_{\mathcal{A}}$). So this implies that to be rational according to Usual Dutch Book Criterion must have probabilistic credences.

This theorem relies on the assumption that $\text{Worlds}_{\mathcal{A}}$ is fixed as c varies, an assumption that is legitimate for agendas that do not depend on the agent's credences. For sentences like π, a choice of the agent's credences affect the truth of π, and therefore the payoff of the bet. So the only loss that the agent should care about is those in the worlds where the agent has the credences she is considering. And betting at some values is basically choosing some credences, since we are making the assumption that an agent must bet with his credences. So for a credal state c, we should not focus on all of $\text{Worlds}_{\mathcal{A}}$, but just on those members which cohere with c. If we are considering a self-ref agenda, this will say that when an agent is considering c, he should only care about his losses in w_c.

So it turns out that the formulation of Dutch books is wrong when we consider such sentences. We will now work through an example of how the agent's losses look when he bets on π.

Example 8.2.3. Consider $\pi \leftrightarrow \neg\mathsf{P}^\ulcorner\pi^\urcorner \geqslant {}^1\!/_2$.

Suppose $c(\pi) \geqslant {}^1\!/_2$. Then the agent will pay $\pounds c(\pi)$ for $(1, \pi)$. But since $c(\pi) \geqslant {}^1\!/_2$, π will be false. So he will lose the bet, and get $\pounds 0$, resulting in a total loss of $\pounds c(\pi)$.

More precisely: Since we have supposed that $c(\pi) \geqslant {}^1\!/_2$, we must be in a world w where $\mathsf{w}(\pi) = 0$. So then

$$\mathsf{Loss}_{(1,\pi)}(c, \mathsf{w}) = c(\pi) - \mathsf{w}(\pi) = c(\pi) \geqslant {}^1\!/_2 > 0$$

And this holds for all ws where the agent has those credences.

Now suppose instead that $c(\pi) \ngeqslant {}^1\!/_2$. Then π must be true. Consider his loss in the bet $(-1, \pi)$. We are then in some w where $\mathsf{w}(\pi) = 1$ and his loss will be:

$$\mathsf{Loss}_{(-1,\pi)}(c, \mathsf{w}) = -c(\pi) - (-\mathsf{w}(\pi)) = -c(\pi) + 1 > 1 - {}^1\!/_2 = {}^1\!/_2 > 0$$

I.e. he will pay $\pounds{-c(\pi)}$ for the bet, and will win $\pounds{-1}$ since then π would be true.

So once the agent's credences are fixed, he will be led to accepting bets which guarantee him a monetary loss.

When we consider self-ref agendas, where a choice of credences fix the truth values of the sentences, the payoff of each bet will depend only on the agent's credences. So once an agent fixes his credences, his loss will be a guaranteed one.

We introduce some notation to refer to this guaranteed loss.

Definition 8.2.4. Suppose $\mathcal{A}$ is a self-ref agenda.

The agent's loss on the set of bets $\mathcal{B}$ is:

$$\mathsf{GuarLoss}_{\mathcal{B}}(c) := \mathsf{Loss}_{\mathcal{B}}(c, \mathsf{w}_c)$$

If $\mathcal{A}$ is a self-ref agenda, and $c \in \mathsf{Creds}_{\mathcal{A}}$, then $\mathcal{B}$ is a *Dutch book* against c if

$$\mathsf{GuarLoss}_{\mathcal{B}}(c) > 0.$$

We now have two definitions of what a Dutch book against c is; the first was given in Definition 8.2.2 and applied to credences defined on agendas which did not refer to probability, the second was given just now and applies to credences defined on self-ref agendas.

As a particular example of $\mathsf{GuarLoss}_{\mathcal{B}}$, consider when $\mathcal{A}$ is a δ-agenda.

Proposition 8.2.5. *For the case where $\mathcal{A}$ is a δ-agenda with $\delta \leftrightarrow \text{`}c \in R\,\text{'}$, we therefore have*[7]

$$\mathsf{GuarLoss}_{\mathcal{B}}(c) = \begin{cases} \mathsf{Loss}_{\mathcal{B}}(c, \mathsf{w}_{\delta}) & c \in R \\ \mathsf{Loss}_{\mathcal{B}}(c, \mathsf{w}_{\neg\delta}) & c \notin R \end{cases}$$

The losses act nicely under extensions of self-ref agendas.

Proposition 8.2.6. *If $\mathcal{A}' \subseteq \mathcal{A}$ are self-ref agendas, and suppose $c \in \mathsf{Creds}_{\mathcal{A}}$ and $c' \in \mathsf{Creds}_{\mathcal{A}'}$ are such that $c = c' \restriction_{\mathcal{A}}$, i.e. for $\varphi \in \mathcal{A}$, $c(\varphi) = c'(\varphi)$. Let $\mathcal{B}$ be a set of bets defined on $\mathcal{A}'$. Then:*

$$\mathsf{GuarLoss}_{\mathcal{B}}(c) = \mathsf{GuarLoss}_{\mathcal{B}}(c')$$

We saw in Example 8.2.3 that whatever credence the agent had in π, there is a bet which would lead him to a guaranteed loss. Sentences which have this feature are *undermining*.

Undermining sentences are the really problematic ones. Ideally if δ were true, one would want to have credences corresponding to w_{δ}. But for the undermining δ, having such a credal state implies that δ is false. And similarly if δ were false, one would want to have credences corresponding to $\mathsf{w}_{\neg\delta}$. But for the undermining δ, this will imply that δ is true. So these are the sentences where one cannot have the *ideal* credences without undermining oneself.

Definition 8.2.7. Suppose δ is such that there is some δ-agenda, and let $\mathcal{A}$ be the minimal such δ-agenda.

We say that δ is *undermining* if

$$\mathsf{w}_{\mathsf{c}_{\delta}}(\delta) = 0$$
$$\text{and } \mathsf{w}_{\mathsf{c}_{\neg\delta}}(\delta) = 1$$

c_{δ} refers to the credal state such that $\mathsf{c}_{\delta}(\delta) = 1$ and $\mathsf{c}_{\delta}(\neg\delta) = 0$.

Example 8.2.8. • A probabilistic liar, π, with $\pi \leftrightarrow \mathsf{P}\ulcorner\pi\urcorner \leqslant 1/2$ is undermining.

---

[7]Therefore:

$$\mathsf{GuarLoss}_{\{(r,\delta),(q,\neg\delta)\}}(c) = \begin{cases} rc(\delta) - r + qc(\neg\delta) & \langle c(\delta), c(\neg\delta)\rangle \in R \\ rc(\delta) + qc(\neg\delta) - q & \langle c(\delta), c(\neg\delta)\rangle \notin R \end{cases}$$

$$|\mathsf{GuarLoss}_{(1,\delta)}(c)| = \begin{cases} |1 - c(\delta)| & c \in R \\ |c(\delta)| & c \notin R \end{cases}$$

$$|\mathsf{GuarLoss}_{(1,\neg\delta)}(c)| = \begin{cases} |c(\neg\delta)| & c \in R \\ |1 - c(\neg\delta)| & c \notin R \end{cases}$$

$$|\mathsf{GuarLoss}_{\{(1,\delta),(1,\neg\delta)\}}(c)| = |c(\delta) + c(\neg\delta) - 1|$$

- A probabilistic truth-teller, η, with $\eta \leftrightarrow \mathsf{P}^{\ulcorner}\eta^{\urcorner} \geqslant 1/2$ is not undermining.

Undermining δs also lead to guaranteed losses. To state this result we first state facts about bets:

Definition 8.2.9. For a set of bets $\mathcal{B} = \{(r_1, \varphi_1), \ldots, (r_n, \varphi_n)\}$, define:

$$-\mathcal{B} := \{(-r_1, \varphi_1), \ldots, (-r_n, \varphi_n)\}$$

Proposition 8.2.10. *For each collection of bets, $\mathcal{B}$,*

$$\mathsf{GuarLoss}_{-\mathcal{B}}(c) = -\mathsf{GuarLoss}_{\mathcal{B}}(c).$$

As a result of this proposition, $\mathsf{GuarLoss}_{\mathcal{B}}(c) > 0$ iff $\mathsf{GuarLoss}_{-\mathcal{B}}(c) < 0$. Therefore, either $\mathcal{B}$ leads to a guaranteed loss, or $-\mathcal{B}$ leads to a guaranteed loss, or both $\mathcal{B}$ and $-\mathcal{B}$ lead to a guaranteed break-even.

Proposition 8.2.11. *The agent's guaranteed loss on whichever of $\mathcal{B}$ or $-\mathcal{B}$ doesn't lead to a gain is $|\mathsf{GuarLoss}_{\mathcal{B}}(c)|$.*
The agent is not led to a guaranteed loss on either $\mathcal{B}$ or $-\mathcal{B}$ iff $\mathsf{GuarLoss}_{\mathcal{B}}(c) = 0$.

So we can now state the result saying that considering some undermining δ leads an agent to a guaranteed loss. This also says that there is a Dutch book against every credal state.

Theorem 8.2.12. *Let $\mathcal{A}$ be a δ-agenda and suppose δ is undermining.*
Then for each $c \in \mathsf{Creds}_{\mathcal{A}}$, there is some $\varphi \in \mathcal{A}$ where one of the bets $(1, \varphi)$ or $(-1, \varphi)$ will lead to a guaranteed loss.

Proof. Suppose c doesn't lead to a loss on any of these unit bets. Then for each $\varphi \in \mathcal{A}$,

$$0 = |\mathsf{GuarLoss}_{(1,\varphi)}(c)| = |\mathsf{Loss}_{(1,\varphi)}(c, \mathsf{w}_c)| = |c(\varphi) - \mathsf{w}_c(\varphi)|$$

so $c(\varphi) = \mathsf{w}_c(\varphi)$ for each $\varphi \in \mathcal{A}$, so in fact $c = \mathsf{w}_c$.
Since $\mathcal{A}$ is a δ-agenda, $\mathsf{w}_c \in \{\mathsf{w}_\delta, \mathsf{w}_{\neg\delta}\}$.

$$\mathsf{w}_c = \mathsf{w}_\delta \implies \mathsf{w}_\delta = \mathsf{w}_{\mathsf{c}_\delta} \implies \mathsf{w}_{\mathsf{c}_\delta}(\delta) = 1$$
$$\mathsf{w}_c = \mathsf{w}_{\neg\delta} \implies \mathsf{w}_{\neg\delta} = \mathsf{w}_{\mathsf{c}_{\neg\delta}} \implies \mathsf{w}_{\mathsf{c}_{\neg\delta}}(\delta) = 0$$

So either $\mathsf{w}_{\mathsf{c}_\delta}(\delta) = 1$ or $\mathsf{w}_{\mathsf{c}_{\neg\delta}}(\delta) = 0$, and therefore δ cannot be undermining. $\qquad\square$

The existence of some agenda that leads to a guaranteed loss was already shown in Caie (2013), namely the agenda $\{\pi, \neg\pi\}$ because of the result described in Example 8.2.3. We have extended that result by showing a range of situations where an agent is lead to a guaranteed loss. Since we have shown that the agent will be to a guaranteed loss on one of the single unit bets, we cannot avoid this problem by restricting our attention to such bets.

The question, which we will attempt to answer in Sections 8.3 and 8.4, then is:

> Do Dutch book considerations lead to any rational constraints, and how should Usual Dutch Book Criterion be modified?

8.3   Failed attempts to modify the criterion

One could of course include an "if possible" constraint so then it is always satisfiable, resulting in the suggestion:

Rationality Criterion 4 (Usual Dutch Book With "If Possible"). *If possible, an agent should have credences c such that agent is not guaranteed a loss on some $\mathcal{B}$.*

In fact we will add an "if possible" clause to each of our considered rational requirements, however we want to do more than that because there are situations where this criterion will then fall silent, but where there are credal states that look better or worse from a Dutch book perspective.

For example in the case of π, having credence $c(\pi) = 0$ will guarantee our agent a loss of £1 in the bet $(-1, \pi)$, whereas having a credal state with $c(\pi) = 0.6$ will guarantee him a loss of £0.60 in the same bet. So all other things being equal, he should prefer a credal state with $c(\pi) = 0.6$ to one where $c(\pi) = 0$. So perhaps we could give an alternative criterion which still leads to constraints in some cases where the agent is guaranteed a loss.

Caie does give a suggestion for how to modify Usual Dutch Book With "If Possible" to do this.

> **LOSS MINIMIZATION** If an agent is guaranteed to value as fair some set of losing bets, then the agent should, if possible, have a credal state that minimizes her possible loss. (Caie, 2013, p. 573)

To formalise this we will need to say what it for an agent to *minimize his possible guaranteed losses.*

In virtue of Proposition 8.2.11 we can attempt to explicate Caie's proposal by saying that an agent should be minimizing the $|\mathsf{GuarLoss}_{\mathcal{B}}(c)|$. The agent's guaranteed loss on a set of bets $\mathcal{B}$ is exactly the same as his guaranteed gain on $-\mathcal{B}$. So by minimizing his possible losses, he also minimizes his possible gains. A risk-seeking agent might take that chance and just hope that the bookie will choose the bets which he will be able to achieve a guaranteed gain from. So in stating Minimize Loss With All $\mathcal{B}$ we are in a sense assuming that our agent is risk-avoiding.

Caie does not explicitly mention which bets to consider, but one proposal might be to minimize $|\mathsf{GuarLoss}_{\mathcal{B}}(c)|$ for *all* collections of bets $\mathcal{B}$. However, we will show that that is the wrong collection of bets to consider as it often falls silent in cases where it shouldn't.

Rationality Criterion 5 (Minimize Loss With All $\mathcal{B}$). *Suppose $\mathcal{A}$ is a self-ref agenda. An agent should, if possible, have credences c such that for each $\mathcal{B}$, the agent minimizes his loss on whichever of $\mathcal{B}$ and $-\mathcal{B}$ do not lead to a gain. I.e. such that $|\mathsf{GuarLoss}_{\mathcal{B}}(c)|$ is minimized.*

We can show that Minimize Loss With All $\mathcal{B}$ is also very often not satisfiable. In fact we can show that it is never satisfied for undermining δ, at least whenever $\{\delta, \neg\delta\}$ is a δ-agenda. So it hasn't achieved much gain, if any, over the original Usual Dutch Book Criterion. This result is Corollary 8.3.3.

Proposition 8.3.1. *Let $\mathcal{A} \supseteq \{\delta, \neg\delta\}$ be a self-ref agenda. If the agent minimizes both $|\mathsf{GuarLoss}_{(1,\delta)}(c)|$ and $|\mathsf{GuarLoss}_{(1,\neg\delta)}(c)|$ then $c(\delta) \in \{0, 1\}$ or $c(\neg\delta) \in \{0, 1\}$.*

Proof. Suppose that b with $b(\delta) \notin \{0,1\}$ minimizes both of these. We have two cases to consider: firstly if $\mathsf{w}_b(\delta) = 1$, we will show that then $b(\neg\delta) = 1$, secondly if $\mathsf{w}_b(\delta) = 0$, we can then show that $b(\neg\delta) = 0$.

Suppose $\mathsf{w}_b(\delta) = 1$. Then $|\mathsf{GuarLoss}_{(1,\delta)}(b)| = |b(\delta) - \mathsf{w}_b(\delta)| = |b(\delta) - 1| > 0$. Let $c_{\langle 1,1 \rangle}$ refer to some credal state in $\mathsf{Creds}_{\mathcal{A}}$ with $c_{\langle 1,1 \rangle}(\delta) = 1$ and $c_{\langle 1,1 \rangle}(\neg\delta) = 1$. Observe that, since b is assumed to minimize $|\mathsf{GuarLoss}_{(1,\delta)}(c)|$, we must have:

$$|\mathsf{GuarLoss}_{(1,\delta)}(c_{\langle 1,1 \rangle})| \geqslant |\mathsf{GuarLoss}_{(1,\delta)}(b)| > 0$$
$$\text{so} \quad 0 < |c_{\langle 1,1 \rangle}(\delta) - \mathsf{w}_{c_{\langle 1,1 \rangle}}(\delta)| = |1 - \mathsf{w}_{c_{\langle 1,1 \rangle}}(\delta)|$$
$$\text{so} \quad \mathsf{w}_{c_{\langle 1,1 \rangle}}(\delta) \neq 1$$
$$\text{so} \quad \mathsf{w}_{c_{\langle 1,1 \rangle}}(\delta) = 0$$

Now, consider $(1, \neg\delta)$.

$$|\mathsf{GuarLoss}_{(1,\neg\delta)}(c_{\langle 1,1 \rangle})| = |c_{\langle 1,1 \rangle}(\neg\delta) - 1| = 0$$

So, since b is assumed to minimize $|\mathsf{GuarLoss}_{(1,\neg\delta)}(c)|$, we have:

$$\text{so} \quad 0 = |\mathsf{GuarLoss}_{(1,\neg\delta)}(b)| = |b(\neg\delta) - 1|$$
$$\text{so} \quad b(\neg\delta) = 1$$

A similar argument holds for $\mathsf{w}_b(\delta) = 0$. Suppose $\mathsf{w}_b(\delta) = 0$. Since $\mathsf{GuarLoss}_{(1,\delta)}(b) = 0$, it must be that $\mathsf{GuarLoss}_{(1,\delta)}(c_{\langle 0,0 \rangle}) > 0$, so $\mathsf{w}_{c_{\langle 0,0 \rangle}}(\delta) = 0$, and therefore $\mathsf{GuarLoss}_{(1,\neg\delta)}(c_{\langle 0,0 \rangle}) = 0$. It must therefore be that $0 = \mathsf{GuarLoss}_{(1,\neg\delta)}(b) = |1 - b(\neg\delta)|$, so $b(\neg\delta) = 1$. $\qquad\square$

Proposition 8.3.2. *Let $\mathcal{A}$ be a self-ref agenda and $c \in \mathsf{Creds}_{\mathcal{A}}$. If c minimizes $|\mathsf{GuarLoss}_{\{(1,\delta),(1,\neg\delta)\}}(c)|$ then $c(\neg\delta) = 1 - c(\delta)$.*

Proof. $|\mathsf{GuarLoss}_{\{(1,\delta),(1,\neg\delta)\}}(c)| = |c(\delta) + c(\neg\delta) - 1|$. This is equal to 0 iff $c(\neg\delta) = 1 - c(\delta)$, so it is minimized at exactly these points. $\qquad\square$

Corollary 8.3.3. *Suppose δ is undermining and $\{\delta, \neg\delta\}$ is a δ-agenda. Then for each $c \in \mathsf{Creds}_{\{\delta,\neg\delta\}}$ there is some*

$$\mathcal{B} \in \{(1,\delta), (1,\neg\delta), \{(1,\delta),(1,\neg\delta)\}\}$$

such that c does not minimize $|\mathsf{GuarLoss}_{\mathcal{B}}(c)|$.

Proof. Suppose b minimizes each of $|\mathsf{GuarLoss}_{(1,\delta)}(b)|$, $|\mathsf{GuarLoss}_{(1,\neg\delta)}(b)|$ and $|\mathsf{GuarLoss}_{\{(1,\delta),(1,\neg\delta)\}}(b)|$. Then by Propositions 8.3.1 and 8.3.2, it must be that $b(\neg\delta) = 1 - b(\delta)$ and either $b(\delta) \in \{0,1\}$ or $b(\neg\delta) \in \{0,1\}$. It must therefore be that $b = \mathsf{w}_\delta$ or $b = \mathsf{w}_{\neg\delta}$. But, we have assumed that δ is under-

mining, we can see that $|\mathsf{GuarLoss}_{(1,\delta)}(b)| = |\mathsf{GuarLoss}_{(1,\neg\delta)}(b)| = 1,$[8] whereas $|\mathsf{GuarLoss}_{(1,\delta)}(c_{\langle 0.5,0.5\rangle})| = |\mathsf{GuarLoss}_{(1,\neg\delta)}(c_{\langle 0.5,0.5\rangle})| = 0.5$. So b does not minimize each of these. $\qquad\square$

Corollary 8.3.5. *Suppose $\{\delta, \neg\delta\}$ is a δ-agenda and δ is undermining. If $\mathcal{A} \supseteq \{\delta, \neg\delta\}$ then for each $c \in \mathsf{Creds}_\mathcal{A}$ there is some $\mathcal{B} \in \{(1, \delta), (1, \neg\delta), \{(1, \delta), (1, \neg\delta)\}\}$ such that c does not minimize $|\mathsf{GuarLoss}_\mathcal{B}(c)|$.*

Proof. Follows directly from Proposition 8.3.2 using Proposition 8.2.6. $\qquad\square$

The first way we consider altering this requirement is to restrict the kinds of bets considered. Caie might have had this kind of restriction in mind as he only explicitly considers the single bets on δ and $\neg\delta$ and not *sets* of bets.

Definition 8.3.6. For a set of bets $\mathcal{B} = \{(r_1, \varphi_1), \ldots, (r_n, \varphi_n)\}$, define:

$$q \cdot \mathcal{B} := \{(q \cdot r_1, \varphi_1), \ldots, (q \cdot r_n, \varphi_n)\}$$

If we are only interested in losses on single bets, we only need to consider the unit bets because of the following result.

Proposition 8.3.7.

$$\mathsf{GuarLoss}_{r\mathcal{B}}(c) = r \cdot \mathsf{GuarLoss}_\mathcal{B}(c)$$

Therefore,

$$|\mathsf{GuarLoss}_{(r,\varphi)}(c)| = |r| \cdot |(1, \varphi)|.$$

Note that we also have:[9]

$$\mathsf{GuarLoss}_{\{(r_1,\varphi_1),\ldots,(r_n,\varphi_n)\}}(c)$$
$$= r_1 \cdot \mathsf{GuarLoss}_{(1,\varphi)}(c) + \ldots + r_n \cdot \mathsf{GuarLoss}_{(1,\varphi_n)}(c).$$

This shows that we only need to consider the $|\mathsf{GuarLoss}_{(r,\varphi)}(c)|$ for $r = 1$, i.e. only consider unit bets, as minimizing this will lead to minimizing the losses on single bets at any stake.

Rationality Criterion 6 (Minimize Guaranteed Losses Unit Bets)**.** *Suppose $\mathcal{A}$ is a self-ref agenda. An agent should if possible have credences c such that agent minimizes each of $|\mathsf{GuarLoss}_{(1,\varphi)}(c)|$ for $\varphi \in \mathcal{A}$.*

---

[8]Because:

Proposition 8.3.4. *Suppose $\mathcal{A} \supseteq \{\delta, \neg\delta\}$ is a self-ref agenda and δ is undermining. Then for $b \in \{\mathsf{w}_\delta, \mathsf{w}_{\neg\delta}\}$ and $\varphi \in \{\delta, \neg\delta\}$, we have*

$$|\mathsf{GuarLoss}_{(1,\varphi)}(b)| = 1$$

Proof.

$$|\mathsf{GuarLoss}_{(1,\delta)}(\mathsf{w}_\delta)| = |\mathsf{w}_\delta(\delta) - \mathsf{w}_{\mathsf{c}_\delta}(\delta)| = |1 - 0| = 1$$
$$|\mathsf{GuarLoss}_{(1,\neg\delta)}(\mathsf{w}_\delta)| = |\mathsf{w}_\delta(\neg\delta) - \mathsf{w}_{\mathsf{c}_\delta}(\neg\delta)| = |0 - 1| = 1$$
$$|\mathsf{GuarLoss}_{(1,\delta)}(\mathsf{w}_{\neg\delta})| = |\mathsf{w}_{\neg\delta}(\delta) - \mathsf{w}_{\mathsf{c}_{\neg\delta}}(\delta)| = |0 - 1| = 1$$
$$|\mathsf{GuarLoss}_{(1,\neg\delta)}(\mathsf{w}_{\neg\delta})| = |\mathsf{w}_{\neg\delta}(\neg\delta) - \mathsf{w}_{\mathsf{c}_{\neg\delta}}(\neg\delta)| = |1 - 0| = 1 \qquad\square$$

[9]But note that $|\mathsf{GuarLoss}_{\{(r_1,\varphi_1),\ldots,(r_n,\varphi_n)\}}(c)|$ is not necessarily equal to $r_1 \cdot |\mathsf{GuarLoss}_{(1,\varphi)}(c)| + \ldots + r_n \cdot |\mathsf{GuarLoss}_{(1,\varphi_n)}(c)|$.

This is satisfiable in more situations. For example it is satisfiable in the case of $\pi \leftrightarrow \neg P^{\ulcorner}\pi^{\urcorner} \geqslant 1/2$, and in fact whenever $\{\delta\}$ is a self-ref agenda.

Example 8.3.8. Consider $\pi \leftrightarrow \neg P^{\ulcorner}\pi^{\urcorner} \geqslant 1/2$ and $\mathcal{A} = \{\pi, \neg\pi\}$.

Observe that

$$|\mathsf{GuarLoss}_{(1,\pi)}(c)| = \begin{cases} |1 - c(\pi)| & c(\pi) \ngeqslant 1/2 \\ |c(\pi)| & c(\pi) \geqslant 1/2 \end{cases}$$

$$|\mathsf{GuarLoss}_{(1,\neg\pi)}(c)| = \begin{cases} |c(\neg\pi)| & c(\pi) \ngeqslant 1/2 \\ |1 - c(\neg\pi)| & c(\pi) \geqslant 1/2 \end{cases}$$

So we can consider the losses geometrically to be the horizontal and vertical distances as in Fig. 8.1.
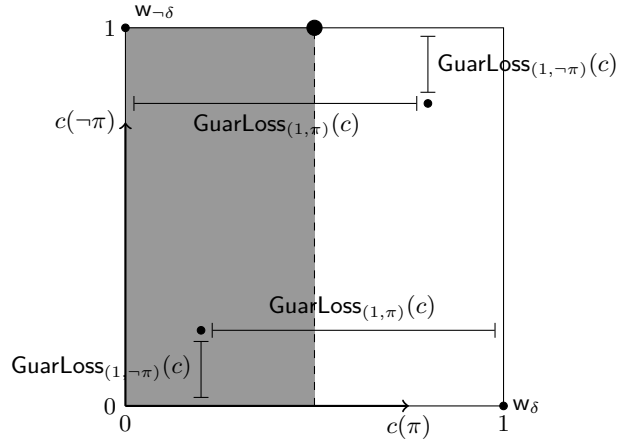
Figure 8.1: A visualisation of $\mathsf{GuarLoss}_{(1,\pi)}(c)$ and $\mathsf{GuarLoss}_{(1,\neg\pi)}(c)$.

It is then clear that the horizontal and vertical distances are both minimized at the point $\langle 0.5, 1 \rangle$.

Working through non-geometrically we have:

If $c(\pi) < 1/2$, then $|\mathsf{GuarLoss}_{(1,\pi)}(c)| = |1 - c(\pi)| > 1/2$.

If $c(\pi) \geqslant 1/2$, then $|\mathsf{GuarLoss}_{(1,\pi)}(c)| = |c(\pi)| \geqslant 1/2$.

So the minimal guaranteed loss on $(1, \pi)$ is $1/2$, which is obtained whenever $c(\pi) = 1/2$.

For $c(\pi) = 1/2$, then $|\mathsf{GuarLoss}_{(1,\neg\pi)}(c)| = |1 - c(\neg\pi)|$, which $= 0$ iff $c(\neg\pi) = 1$.

So by having credal state $c(\pi) = 1/2$ and $c(\neg\pi) = 1$ the agent can minimize his guaranteed losses on both $(1, \pi)$ and $(1, \neg\pi)$.

However, due to Proposition 8.3.1, to be rational one must have extremal credences. This is an undesirable feature.

We also know that in some situations this criterion falls silent, for example:

Proposition 8.3.9. *Consider*

$$\delta \leftrightarrow (P^{\ulcorner}\delta^{\urcorner} \leqslant 0.5 \vee (P^{\ulcorner}\delta^{\urcorner} \leqslant 0.55 \wedge P^{\ulcorner}\neg\delta^{\urcorner} \geqslant 0.2)).$$

There is no c that minimizes both $|\mathsf{GuarLoss}_{(1,\delta)}(c)|$ and $|\mathsf{GuarLoss}_{(1,\neg\delta)}(c)|$.

206

Proof. We represent this situation diagrammatically in Fig. 8.2.
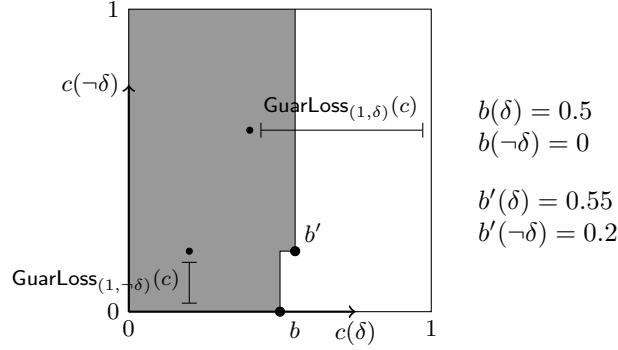
Figure 8.2: Example of δ where no credal state minimizes both $\mathsf{GuarLoss}_{(1,\delta)}(c)$ and $\mathsf{GuarLoss}_{(1,\neg\delta)}(c)$.

Then

- $|\mathsf{GuarLoss}_{(1,\delta)}(b)| = 0.5$
- $|\mathsf{GuarLoss}_{(1,\delta)}(b')| = 0.45$
- $|\mathsf{GuarLoss}_{(1,\neg\delta)}(b)| = 0$
- $|\mathsf{GuarLoss}_{(1,\neg\delta)}(b')| = 0.2$

So neither b nor b' minimizes both.

$|\mathsf{GuarLoss}_{(1,\delta)}(c)|$ can, as in Example 8.3.8, be visualised as the horizontal distance between the credal state and the appropriate vertical axis (depending on whether $\mathsf{w}_c(\delta) = 1$ or not). So the minimal such distance is obtained at credal states $\langle 0.55, y \rangle$ with $y \geqslant 0.2$.

But for such credal states $|\mathsf{GuarLoss}_{(1,\neg\delta)}(c)| \geqslant 0.2$, whereas $|\mathsf{GuarLoss}_{(1,\neg\delta)}(b)| = 0$. $\qquad\square$

The way that we propose to instead weaken Minimize Loss With All $\mathcal{B}$ is to ask that the agent should minimize his *overall* guaranteed loss (over the bets that do not guarantee a gain). This will play a role when the agent cannot simultaneously minimize each of the guaranteed losses, as for example in Proposition 8.3.9.

8.4 The proposal – minimize the overall guaranteed loss

We now need to characterise what it is for an agent to minimize his *overall guaranteed loss*.

If we only consider single bets and not collections of bets, we just need to evaluate the agent's guaranteed losses on the unit bets, i.e. $|\mathsf{GuarLoss}_{(1,\varphi)}(c)|$ for $\varphi \in \mathcal{A}$. This is because $\mathsf{GuarLoss}_{(r,\varphi)}(c) = r \cdot \mathsf{GuarLoss}_{(1,\varphi)}(c)$.

There are two natural ways to measure the agents overall losses on the unit bets. The first is to minimize his *average* losses, and the second is his *maximum* losses. This leads us to the following two possible criteria:

Rationality Criterion 7 (Minimize Average-Unit-Guaranteed-Loss). *Suppose $\mathcal{A}$ is a (finite) self-ref agenda.*

An agent should, if possible, have credences c such that he minimizes his average-unit-guaranteed-loss, which is:

$$\frac{\sum_{\varphi \in \mathcal{A}} |\mathsf{GuarLoss}_{(1,\varphi)}(c)|}{|\mathcal{A}|}$$

Rationality Criterion 8 (Minimize Maximum-Unit-Guaranteed-Loss). [10] *Suppose $\mathcal{A}$ is a (finite) self-ref agenda.*

An agent should, if possible, have credences that minimize his maximum-unit-guaranteed-loss, *which is:*

$$\max\{|\mathsf{GuarLoss}_{(1,\varphi)}(c)| \mid \varphi \in \mathcal{A}\}$$

Consider again the example from Proposition 8.3.9.

Example 8.4.1. Consider

$$\delta \leftrightarrow (\mathsf{P}^{\ulcorner}\delta^{\urcorner} \leqslant 0.5 \vee (\mathsf{P}^{\ulcorner}\delta^{\urcorner} \leqslant 0.55 \wedge \mathsf{P}^{\ulcorner}\neg\delta^{\urcorner} \geqslant 0.2)).$$

And the agenda $\{\delta, \neg\delta\}$. Let $b = \langle 0.5, 0 \rangle$ and $b' = \langle 0.55, 0.2 \rangle$ as in Fig. 8.2.

We have:

- $|\mathsf{GuarLoss}_{(1,\delta)}(b)| = 0.5$

- $|\mathsf{GuarLoss}_{(1,\delta)}(b')| = 0.45$

- $|\mathsf{GuarLoss}_{(1,\neg\delta)}(b)| = 0$

- $|\mathsf{GuarLoss}_{(1,\neg\delta)}(b')| = 0.2$

So neither b nor b' minimizes both guaranteed losses and also neither does any other credal state.

The agent's average-unit-guaranteed-loss is:

- At b:
$$\frac{0.5 + 0}{2} = 0.25$$

- At b':
$$\frac{0.45 + 0.2}{2} = 0.325$$

So Minimize Average-Unit-Guaranteed-Loss says that b is better than b', and in fact it will lead to b as being required by rationality because there is no other credal state that has lower average losses.

The agent's maximum-unit-guaranteed-loss is:

- At b:
$$\max\{0.5, 0\} = 0.5$$

- At b':
$$\max\{0.45, 0.2\} = 0.45$$

---

[10]This suggestion was made to me by Patrick LaVictoire.

8.4 The proposal – minimize the overall guaranteed loss

So Minimize Maximum-Unit-Guaranteed-Loss says that b' is better than b, and in fact it will lead to b' as being required by rationality because there is no other credal state that has lower maximum losses.

The problem with Minimize Maximum-Unit-Guaranteed-Loss is that it does not require the agent to minimize both of $|\mathsf{GuarLoss}_{(1,\delta)}(c)|$ and $|\mathsf{GuarLoss}_{(1,\neg\delta)}(c)|$ if he can do so. So it does not imply Minimize Guaranteed Losses Unit Bets.
For example:

Example 8.4.2. Consider again

$$\pi \leftrightarrow \neg\mathsf{P}^\ulcorner\pi^\urcorner \geqslant {}^1\!/_2$$

and $\mathcal{A} = \{\pi, \neg\pi\}$.
As in Example 8.3.8, $|\mathsf{GuarLoss}_{(1,\pi)}(c)| \geqslant {}^1\!/_2$ and $= {}^1\!/_2$ iff $c(\pi) = {}^1\!/_2$.
Now, for the credal states with $c(\pi) = {}^1\!/_2$, π is true, so $|\mathsf{GuarLoss}_{(1,\neg\pi)}(c)| = |c(\neg\pi)|$ which is minimized (and $= 0$) at $c(\neg\pi) = 1$.
So the credal state $\langle {}^1\!/_2, 1 \rangle$ minimizes the agent's losses on bets on π and $\neg\pi$. However, for all the credal states $\langle {}^1\!/_2, r \rangle$ with $r \in [{}^1\!/_2, {}^3\!/_2]$, The agent's maximum-unit-guaranteed-loss is:

$$\max\{|1 - r|, {}^1\!/_2\} = {}^1\!/_2$$

So this measure sees each of these credal states as equally good. It therefore would not lead to the requirement to have credal state $\langle {}^1\!/_2, 1 \rangle$ but would count as rationally permissible all these credal states. If our agent is in fact keen to minimize his overall losses, this would seem to be a bad way to do it as he leaves himself open to more possible guaranteed losses. See Fig. 8.3.
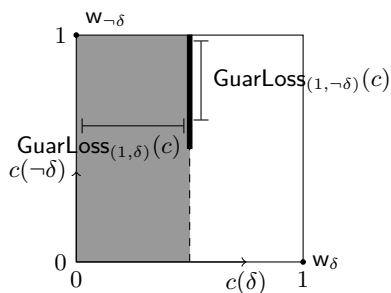
Figure 8.3: The credal states that are permissible according to Minimize Maximum-Unit-Guaranteed-Loss

If one is interested in minimizing overall guaranteed loss, this then seems to be a bad measure. We therefore suggest that we should focus on minimizing the expected loss, as given by Minimize Average-Unit-Guaranteed-Loss.
In this criterion we focused only on single bets and not on collections of bets. That allows us simplicity, but perhaps would lead to undesirable consequences as it might deem one credal state rationally required which is good when only single bets are considered but very bad according to *sets of* bets. If an agent is up for taking sets of bets he should also try to minimize his losses on these.

In Section 8.A we have presented some suggestions for how one could minimize losses also on collections of bets, but we will not study these further.

This kind of question of how to minimize his overall guaranteed loss may connect to the question of measuring how incoherent a credence function must be by determining what Dutch book losses he is susceptible to. This is a question studied by Schervish et al. (2000, 2002, 2003) and discussed in De Bona and Finger (2014); Staffel (2015). Further analysis on the connection between these is left for future work.

8.5 The connection to SelfInacc

We first show that in fact the Dutch book criterion that we have suggested is a version of the accuracy criterion using particular inaccuracy measures. This is an interesting connection.

The definition in Minimize Average-Unit-Guaranteed-Loss means that we can consider his overall loss as a measure of inaccuracy of the agent's credences in terms of the distance from world where he has those credences.

Definition 8.5.1. The *absolute value distance,* AbsValDist is

$$\mathsf{AbsValDist}(\langle x_1, \ldots, x_n \rangle, \langle y_1, \ldots, y_n \rangle) := \frac{\sum_{i=1,\ldots,n} |x_i - y_i|}{n}$$

We use the same name to denote the inaccuracy measure, so for $\mathcal{A} = \{\varphi_1, \ldots, \varphi_n\}$ and $c \in \mathsf{Creds}_{\mathcal{A}}$,

$$\mathsf{AbsValDist}(c, \mathsf{w}) := \mathsf{AbsValDist}(\langle c(\varphi_1), \ldots, c(\varphi_n) \rangle, \langle \mathsf{w}(\varphi_1), \ldots, \mathsf{w}(\varphi_n) \rangle)$$
$$= \frac{\sum_{i=1,\ldots,n} |c(\varphi_i) - \mathsf{w}_c(\varphi_i)|}{n}$$

The un-normalized variants have also been called the taxicab distance, ℓ^1-distance and Manhattan distance.

Proposition 8.5.2.
$$\frac{\sum_{\varphi \in \mathcal{A}} |\mathsf{GuarLoss}_{(1,\varphi)}(c)|}{|\mathcal{A}|} = \frac{\sum_{\varphi \in \mathcal{A}} |c(\varphi) - \mathsf{w}_c(\varphi)|}{|\mathcal{A}|}$$
$$= \mathsf{AbsValDist}(c, \mathsf{w}_c)$$
$$= \mathsf{SelfInacc}^{\mathsf{AbsValDist}}(c)$$

Therefore the rationality constraint from Minimize Average-Unit-Guaranteed-Loss is the same as Minimize Self-Inaccuracy with the inaccuracy measure AbsValDist.

This means that one can consider this as a version of Minimize Self-Inaccuracy. So for self-ref agendas, minimizing overall Dutch book loss corresponds to minimizing self-inaccuracy. AbsValDist is not a proper inaccuracy measure and so it generally taken to be an illegitimate inaccuracy measure. It cannot be used for the traditional accuracy argument because in the traditional accuracy setup, some non-probabilistic credal states will not be dominated. But it does have some prima facie intuitive appeal and it has been argued for in Maher (2002).

AbsValDist satisfies TruthDirectedness and Normality, so we have that the results from Section 7.2 still hold, so it leads to the failure of probabilism, rational introspection and logical omniscience and leads to the requirement to explicitly restrict credences to being in the unit interval.

There is also an interesting connection between Minimize Self-Inaccuracy and Minimize Maximum-Unit-Guaranteed-Loss.

Proposition 8.5.3.

$$\max\{\mathsf{GuarLoss}_{(1,\varphi)}(c) \mid \varphi \in \mathcal{A}\} = \mathsf{SelfInacc}^{\ell^\infty}(c)$$

Therefore the rationality constraint from Minimize Maximum-Unit-Guaranteed-Loss is the same as Minimize Self-Inaccuracy with respect to inaccuracy measure ℓ^∞, *where*

$$\ell^\infty(\langle x_1, \ldots, x_n \rangle, \langle y_1, \ldots, y_n \rangle) := \max\{|x_i - y_i| \mid i \in \{1, \ldots, n\}\}.$$

The other suggestions in Section 8.A will lead to similar correspondences with other $\mathcal{I}$, but it is not yet clear that the $\mathcal{I}$ corresponding to these will satisfy any nice properties.

In fact for any criterion that assigns to each possible credal state some single value of utility, or in fact we consider the *dis*utility, called D, for example the expected loss on bets, or maximum loss on bets, we can understand the criterion of minimizing D as to be the same as minimizing $\mathsf{SelfInacc}^{\mathcal{I}_D}$ for $\mathcal{I}_D$ defined by:

$$\mathcal{I}_D(c, \mathsf{w}_c) := D(c)$$

What is particularly interesting about the criteria that we consider is that they correspond to *natural* inaccuracy measures.

In Section 7.2.5 we showed that Minimize Self-Inaccuracy was inaccuracy measure dependent, so we see that if we choose an alternative inaccuracy measure, like, e.g., the Brier score or the logarithmic score, (self-)accuracy and Dutch book considerations lead to different rationally required credences.

8.6 Don't bet with your credences

A natural response to the challenges faced by the Dutch book criterion is to drop the assumption that the agent should bet with his credences. This is an example of where even an agent who values money positively, in a linear way and does not value anything else at all, positively or negatively, should sometimes *not* sanction as fair monetary bets at odds matching his degrees of belief.

This is an assumption that could not be dropped if we take the traditional behaviourist perspective that all credences are defined or characterised by an agent's betting behaviour: where what it *means* for an agent to have credence ½ in φ is that he will pay anything up to £0.50 for $(1, \varphi)$. However, we do not assume that that is all there is to an agent's credences.

Let's now consider what happens if we drop the assumption that the agent bets with his credences.

8.6.1 How degrees of belief determine the fair betting odds

There are still connections between the agent's degrees of belief and her rational betting behaviour, but the fair betting odds may not be equal to the credences in cases of self-ref agendas.

For example: If an agent currently has $b(\pi) = 0$, then π would be true, so if we are assuming the agent is introspective and aware of this, he should be willing to pay anything up to £1 for $(1, \pi)$. So

$$\mathsf{FairBettingOdds}_b(\pi) = \mathsf{w}_b(\pi)$$

More generally, if we take a self-ref agenda $\mathcal{A}$, and $b \in \mathsf{Creds}_{\mathcal{A}}$ we have

$$\mathsf{FairBettingOdds}_b(\varphi) = \mathsf{w}_b(\varphi).$$

And so when offered a collection of bets, the agent should be willing to pay anything up to:[11]

$$\mathsf{FairBettingOdds}_b(\{(r_1, \varphi_1), \ldots, (r_n, \varphi_n)\}) = \sum_{i=1,\ldots,n} r_i \cdot \mathsf{w}_b(\varphi_i).$$

This is because that is exactly the amount that the agent knows he will win by taking this collection of bets so he should be willing to pay anything up to this value to buy that collection of bets.

We would like to provide a general account of how an agent's fair betting odds should match his credences. This special case where $\mathcal{A}$ is a self-ref agenda is given by $\mathsf{FairBettingOdds}_b(\varphi) = \mathsf{w}_b(\varphi)$. The special case where $\mathcal{A}$ doesn't refer to P, we should have

$$\mathsf{FairBettingOdds}_b(\varphi) = b(\varphi)$$

i.e. the agent should bet in accordance with his credences. But what about for $Heads \vee \pi$? Perhaps

$$\mathsf{FairBettingOdds}_b(Heads \vee \pi) = \begin{cases} b(Heads) & b(\pi) \geqslant 1/2 \\ 1 & \text{otherwise} \end{cases}.$$

It is left to future work to see whether this is the appropriate answer and how this might be generalised.

8.6.2 Moving to the credal state corresponding to the fair betting odds?

So what we have been suggesting in this section is: instead of betting with his current degrees of belief, the agent should instead bet at the fair betting odds. The fair betting odds correspond to some credal state which will be different to his current degrees of belief. Insofar as the pay-off of the bets depend on the agent's credences, they depend on his current degrees of belief instead of the function which corresponds to his betting prices. In this way he can avoid

---

[11]Note that we are abusing terminology because we are using $\mathsf{FairBettingOdds}_b()$ for both sentences and collections of bets. This should not cause confusion because we can say that what we mean by $\mathsf{FairBettingOdds}_b(\varphi)$ is $\mathsf{FairBettingOdds}_b((1, \varphi))$.

being led to a guaranteed loss. For the case of self-ref agendas and introspective agents, if the agent is in credal state b, this then says that he should bet at the odds determined by w_b.[12]

A further step might be to ask the agent to shift his beliefs in accordance with the appropriate fair betting prices. So if he currently in credal state b, then he will want to bet in accordance with credal state w_b, i.e. assign degree of belief 1 or 0 to each sentence depending on whether it is true or not if b is the interpretation of P. However, if he does shift his beliefs in this way, once he has shifted he will again wish to shift: no credal state is *stable*. We showed this in Theorem 8.2.12 by showing that for every credal state (at least those which assigns a numerical degree of belief to some undermining sentence, for example π), there will be some bet where the credences which he should bet with are different from his current credences. So instead we suggest that he should perhaps be content with the situation and just bet with some credal state that is not his own one.

Appendix 8.A   Options for Dutch book criteria

We here present some alternatives to Minimize Average-Unit-Guaranteed-Loss that also consider *sets of* bets.

Our first suggestion is that the agent should minimize his expected losses on the bets of the form $\{(r, \delta), (1 - r, \neg\delta)\}$ with $r \in [0, 1]^2$. These are the bets where there is £1 at stake.

Rationality Criterion 9. *Suppose $\mathcal{A}$ is a self-ref agenda. An agent should have credences that minimize*

$$\frac{1}{2} \cdot \left(\begin{array}{l} \int_0^1 |\mathsf{GuarLoss}_{\{(r,\delta),(1-r,\neg\delta)\}}(c)| \, \mathrm{d}r \\ + \quad \int_0^1 |\mathsf{GuarLoss}_{\{(-r,\delta),(1-r,\neg\delta)\}}(c)| \, \mathrm{d}r \end{array} \right)$$

Another possible suggestion is to focus on the bets $(1, \delta)$, $(1, \neg\delta)$, $\{(1, \delta), (1, \neg\delta)\}$ and $\{(1, \delta), -(1, \neg\delta)\}$.

Rationality Criterion 10. *Suppose $\mathcal{A}$ is a self-ref agenda. An agent should have credences that minimize*

$$\frac{1}{4} \left(\begin{array}{l} |\mathsf{GuarLoss}_{(1,\delta)}(c)| + |\mathsf{GuarLoss}_{(1,\neg\delta)}(c)| \\ + \quad |\mathsf{GuarLoss}_{\{(1,\delta),(1,\neg\delta)\}}(c)| + |\mathsf{GuarLoss}_{\{(1,\delta),-(1,\neg\delta)\}}(c)| \end{array} \right)$$

Alternatively we could consider not minimizing the *expected*, or average, guaranteed loss but instead minimizing the *maximum* guaranteed loss, so more similar now to Rationality Criterion 8. We would have to restrict to bets of a certain "size" to ensure that the guaranteed loss is bounded in order to obtain some rational requirement.

---

[12]We can now observe a connection between this don't-bet-with-your-credences analysis and the accuracy considerations. Our proposal when we considered the accuracy criterion was that was that of 2b from Section 7.3 where (for introspective agents and self-ref agendas) the inaccuracy should be evaluated as $\mathsf{Est}_b(\mathcal{I}(c, \cdot))$, which for any inaccuracy measure that is truth-directed will also be minimized at the credences corresponding to w_b. Perhaps this connection could also be generalised, also using the observations in Section 8.5, but that is left for future work.

Rationality Criterion 11. *An agent should have credences that minimize*

$$\max\{\mathsf{GuarLoss}_{\{(r,\delta),(q,\neg\delta)\}}(c) \mid |r| + |q| \leqslant 1\}$$

A bet of the form $\{(r,\delta),((1-r),\neg\delta)\}$ gives a weighted average of the bets $(1,\delta)$ and $(1,\neg\delta)$. However, the guaranteed loss on the set of bets is not a weighted average of the guaranteed loss on the single bets. As a result, the rational requirements in Minimize Average-Unit-Guaranteed-Loss and may differ.[13]

However, the fact that the bets of the form $\{(r,\delta),(1-r,\neg\delta)\}$ are weighted averages of the unit bets, this suggests that the unit bets are the primitive ones. The motivations between all these three suggestions are the same so we should focus on the one where we consider just unit bets. The result of considering either Rationality Criterion 9 or 10 is that the expected loss is pulled towards favouring the probabilistic credences.

These options should be further studied but we will not do that in this thesis.

---

[13]If it were the case that $|\mathsf{GuarLoss}_{\{(r,\delta),(q,\neg\delta)\}}(c)| = r \cdot |\mathsf{GuarLoss}_{(1,\delta)}(c)| + q \cdot |\mathsf{GuarLoss}_{(1,\neg\delta)}(c)|$ then one would have that

$$\int_0^1 |\mathsf{GuarLoss}_{\{(r,\delta),(1-r,\neg\delta)\}}(c)| \, \mathrm{d}r = \frac{|\mathsf{GuarLoss}_{(1,\delta)}(c)| + |\mathsf{GuarLoss}_{(1,\neg\delta)}(c)|}{2}$$

and therefore that Minimize Average-Unit-Guaranteed-Loss and in fact give the same rational requirements.

Chapter 9

Conclusions

In this thesis we have been studying frameworks which have sentences that can talk about their own probabilities. We have seen that these face a number of challenges.

In Part I we focused on developing semantics and theories for this language. We developed a number of different semantics, which each have their advantages and disadvantages. We have therefore not provided a definitive answer to the question of which sentences are true or not, i.e. which is the correct semantics, but have presented a number of interesting options.

In Chapters 3 and 4 we developed a Kripke-style semantics that worked over probabilistic modal structures in a similar manner to that developed in Halbach and Welch (2009). This allows for fixed points which are in a sense *stable* or *look best from their own perspective*. This informal way of talking is hopefully made precise by saying that these fixed points minimize the estimated inaccuracy from their own perspective. This was considered in Section 7.3.4.

A feature of the semantics from Chapter 3, which uses a strong Kleene evaluation scheme, is that we can see it as assigning sentences intervals as probability values instead of single numbers. Certain axioms of probability have to be dropped, for example $\mathsf{P}_{=1}\ulcorner\lambda\vee\neg\lambda\urcorner$ is not satisfied in the construction. This means that the intervals of probabilities are not the same as those studied the imprecise probability literature. The intervals can instead be seen as non-classical probabilities over (logics arising from) strong Kleene evaluation schemes. A advantage of using this evaluation scheme is that it is compositional and we are able to obtain an axiomatisation which carries a completeness aspect to it once one accepts the ω-rule to fix the standard model of arithmetic.

We obtain imprecise probabilities in Chapter 4, where we consider a supervaluational variant of the Kripke style semantics. In fact there we did not then do something directly analogous to that in Chapter 3 which would have also resulted in assigning intervals of probabilities to sentences.[1] Instead we worked with understanding an agent's credal state as a set of probability functions. This is a variant of imprecise probabilities that has been well studied and it is more general version than just considering intervals. A very nice feature of working with this (possibly non-convex) sets-of-probabilities version of imprecise probabilities is that we obtain the result that in the relevant fixed points every

---

[1]That would have still been different to Chapter 3 because for example $\lambda\vee\neg\lambda$ would have been assigned a precise probability 1.

member of the credal state looks best from some (possibly different) member's perspective. We again suggest that this can be made precise using minimizing estimated inaccuracy considerations, where estimated inaccuracy is calculated by deferring to the probabilistic modal structures, though the exact details and interpretation of this need to be further studied in future work.

In Chapter 5 we considered a revision semantics instead of a Kripke-style semantics. A major advantage of these semantics is that one retains classical logic and traditional probability theory but the price to pay is that one obtains a transfinite sequence of interpretations and identifying any particular interpretation as "correct" is problematic. In the chapter we paid careful attention to finding limit stages that can themselves be used as good models for probability and truth. This is a focus that is not usually present when revision theories are studied. This lead to limit stages where the probability function satisfies the usual (finitely-additive) probability axioms and the truth predicate is maximally consistent; these are features which are not obtained in the usual revision construction. In this chapter we suggested a number of different revision sequences. The style of sequence considered in Section 5.2 used the idea of relative frequencies in the revision sequence up to that stage to define the next interpretation of probability, extending ideas from Leitgeb (2012). This results in an interesting construction that is best understood as providing something like *semantic probabilities* but isn't suitable for providing a semantics for modelling subjective probabilities. We therefore considered, in Section 5.3, using probabilistic modal structures to develop a revision sequence, which allows for this subjective interpretation of the probability notion. In giving this we provided a few options for how to characterise the limit stage. To allow Probabilistic Convention T to be satisfied we gave a limit definition which we showed to be satisfiable using generalised Banach limits. However the revision sequence that we favour rejects Probabilistic Convention T and instead requires the limit stage to sum up the results from the previous stages in a strong way.

One reason to develop a semantics is to better understand these languages. In all the semantics we developed using probabilistic modal structures, we can then see that certain principles may need to be reformulated. For example we noted that if a principle of introspection is instead expressed using a truth predicate to do the job of quotation and disquotation we obtain a principle which is satisfied in exactly the probabilistic modal structures that are introspective (which is exactly those that satisfy the usual introspection principle in the operator language). This style of reformulation was suggested in Stern (2014a,b) for the case of (all-or-nothing) modalities.

In Part II we focused on the subjective interpretation of probability and considered how arguments for rationality requirements apply in such a framework. For self-referential sentences, a choice of the agent's credences will affect which worlds are possible. Caie (2013) argued that the accuracy and Dutch book arguments should be modified because the agent should only care about her inaccuracy or payoffs in the world which could be actual if she adopted the considered credences. We considered the accuracy argument in Chapter 7 and the Dutch book argument in Chapter 8. Both these accuracy and Dutch book criteria mean that an agent is rationally required to be probabilistically incoherent, have negative credences and to fail to assign the same credence to logically equivalent sentences. We also showed that that accuracy criterion depends on how inaccuracy is measured and that it differs from the Dutch book criterion (at

least when the inaccuracy measure is not AbsValDist). We end up rejecting the proposed modification of the accuracy criterion and the Dutch book criterion. In Section 7.3, we reconsidered the accuracy criterion and instead suggested that the agent should consider the estimation of how accurate her credences are, from the perspective of her current credences, where the estimation is taken using the probabilistic modal structure. We also discussed how to generalise this version of the accuracy criterion and presented ideas suggesting that it connects to the semantics developed in Part I. This can then provide some formal meaning to the discussion in Part I that fixed points *look best from their own perspectives.* In Section 8.6 we suggested that this is a case where an agent should not bet with his credences.

There are still many open questions that this thesis has left, but we hope to have given some clarity to the question of how languages with self-referential probabilities may work.

List of Definitions

Bibliography

David S. Ahn. Hierarchies of ambiguous beliefs. *Journal of Economic Theory*, 136(1):286–301, 2007.

Charalambos D Aliprantis and Kim C. Border. *Infinite dimensional analysis*, volume 32006. Springer, 1999.

Robert J. Aumann. Interactive epistemology II: Probability. *International Journal of Game Theory*, 28(3):301–314, 1999.

Fahiem Bacchus. Lp, a logic for representing and reasoning with statistical knowledge. *Computational Intelligence*, 6(4):209–231, 1990.

Darren Bradley and Hannes Leitgeb. When betting odds and credences come apart: More worries for dutch book arguments. *Analysis*, pages 119–127, 2006.

Seamus Bradley. Imprecise probabilities. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer 2015 edition, 2015.

Michael Caie. *Paradox and Belief*. PhD thesis, University of California, Berkeley, 2011.

Michael Caie. Rational probabilistic incoherence. *Philosophical Review*, 122(4): 527–575, 2013.

Michael Caie. Calibration and probabilism. *Ergo*, 1:13–38, 2014.

Catrin Campbell-Moore. How to express self-referential probability. A kripkean proposal. *The Review of Symbolic Logic*, 8:680–704, 12 2015a. ISSN 1755-0211. doi: 10.1017/S1755020315000118. URL http://journals.cambridge.org/article_S1755020315000118.

Catrin Campbell-Moore. Rational probabilistic incoherence? A reply to Michael Caie. *Philosophical Review*, 124(3):393–406, 2015b.

Jennifer Carr. Epistemic utility theory and the aim of belief. Accessed Oct 31. 2014., ms.

Chen Chung Chang and H Jerome Keisler. *Model theory*. Elsevier, 1990.

David Christensen. Clever bookies and coherent beliefs. *The Philosophical Review*, pages 229–247, 1991.

Paul Christiano, Eliezer Yudkowsky, Marcello Herresho, and Mihaly Barasz. Definability of truth in probabilistic logic, early draft. Accessed March 28, 2013, ms.

Glauber De Bona and Marcelo Finger. Notes on measuring inconsistency in probabilistic logic. Technical report, Technical report RT-MAC-2014-02, Department of Computer Science, IME/USP. http://www. ime. usp. br/~ mfinger/www-home/papers/DBF2014-reltec. pdf, 2014.

Andy Egan and Adam Elga. I can't believe i'm stupid. *Philosophical Perspectives*, 19(1):77–93, 2005.

Daniel Ellsberg. Risk, ambiguity, and the savage axioms. *The quarterly journal of economics*, pages 643–669, 1961.

Ronald Fagin, Joseph Y Halpern, and Nimrod Megiddo. A logic for reasoning about probabilities. *Information and computation*, 87(1):78–128, 1990.

Haim Gaifman. A theory of higher order probabilities. In *Causation, chance and credence*, pages 191–219. Springer, 1988.

Haim Gaifman and Marc Snir. Probabilities over rich languages, testing and randomness. *The journal of symbolic logic*, 47(03):495–548, 1982.

Robert Goldblatt. The countable Henkin principle. In *The Life and Work of Leon Henkin*, pages 179–201. Springer, 2014.

Hilary Greaves. Epistemic decision theory. *Mind*, page fzt090, 2013.

Anil Gupta. Truth and paradox. *Journal of philosophical logic*, 11(1):1–60, 1982.

Anil Gupta and Nuel Belnap. The revision theory of truth. *MIT Press, Cambridge*, 1(99):3, 1993.

Alan Hájek. Dutch book arguments. *The handbook of rational and social choice*, pages 173–196, 2008.

Volker Halbach. *Axiomatic theories of truth*. Cambridge Univ Pr, 2011.

Volker Halbach. *Axiomatic Theories of Truth (Revised Edition)*. Cambridge University Press, 2014. ISBN 9781107424425.

Volker Halbach and Philip Welch. Necessities and necessary truths: A prolegomenon to the use of modal logic in the analysis of intensional notions. *Mind*, 118(469):71–100, 2009.

Volker Halbach, Hannes Leitgeb, and Philip Welch. Possible-worlds semantics for modal notions conceived as predicates. *Journal of Philosophical Logic*, 32: 179–222, 2003.

Aviad Heifetz and Philippe Mongin. Probability logic for type spaces. *Games and economic behavior*, 35(1):31–53, 2001.

Pavel Janda. Measuring inaccuracy of uncertain doxastic states in many-valued logical systems. *Journal of Applied Logic*, 14:95–112, 2016.

BIBLIOGRAPHY

James M. Joyce. A Nonpragmatic Vindication of Probabilism. *Philosophy of Science*, 65:575–603, 1998.

James M. Joyce. Accuracy and coherence: Prospects for an alethic epistemology of partial belief. In *Degrees of belief*, pages 263–297. Springer, 2009.

Jason Konek and Ben Levinstein. The foundations of epistemic decision theory. Accessed Oct 31, 2014., ms.

Saul Kripke. Outline of a theory of truth. *The journal of philosophy*, 72(19): 690–716, 1975.

H. Leitgeb. On the probabilistic convention T. *The Review of Symbolic Logic*, 1(02):218–224, 2008.

Hannes Leitgeb. From type-free truth to type-free probability. In Greg Restall and Gillian Russel, editors, *New Waves in Philosophical Logic edited by Restall and Russell*, pages 84–94. Palgrave Macmillan, 2012.

Hannes Leitgeb and Richard Pettigrew. An objective justification of bayesianism I: Measuring inaccuracy. *Philosophy of Science*, 77(2):201–235, 2010.

David Lewis. A subjectivist's guide to objective chance. *Studies in inductive logic and probability*, 2:263–293, 1980.

David Lewis. Humean supervenience debugged. *Mind*, pages 473–490, 1994.

Patrick Maher. *Betting on theories*. Cambridge University Press, 1993.

Patrick Maher. Joyce's argument for probabilism. *Philosophy of Science*, 69(1): 73–81, 2002.

Angelo Margaris. *First order mathematical logic*. Courier Corporation, 1990.

Conor Mayo-Wilson and Gregory Wheeler. Scoring imprecise credences. *Forthcoming in Philosophy and Phenomenological Research*, ta.

Vann McGee. How truthlike can a predicate be? A negative result. *Journal of Philosophical Logic*, 14(4):399–410, 1985.

Martin Meier. Finitely additive beliefs and universal type spaces. *The Annals of Probability*, 34(1):386–422, 2006.

David Miller. A paradox of information. *British Journal for the Philosophy of Science*, pages 59–61, 1966.

Yiannis N. Moschovakis. *Elementary induction on abstract structures*. North-Holland, New York, 1974.

Zoran Ognjanović and Miodrag Rašković. A logic with higher order probabilities. *Publications de l'Institut Mathématique. Nouvelle Série*, 60:1–4, 1996.

Jeff B. Paris. A note on the dutch book method. In *Proceedings of the Second International Symposium on Imprecise Probabilities and their Applications, ISIPTA*, pages 301–306, 2001.

Richard Pettigrew. Epistemic utility arguments for probabilism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2011 edition, 2011.

Richard Pettigrew. Accuracy and the laws of credence. Forthcoming with Oxford University Press, ms.

Frank P. Ramsey. Truth and probability (1926). *The foundations of mathematics and other logical essays*, pages 156–198, 1931.

Jan-Willem Romeijn. Conditioning and interpretation shifts. *Studia Logica*, 100 (3):583–606, 2012.

Mark J Schervish, Teddy Seidenfeld, and Joseph B. Kadane. How sets of coherent probabilities may serve as models for degrees of incoherence. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 8(03):347–355, 2000.

Mark J Schervish, Teddy Seidenfeld, and Joseph B. Kadane. Measuring incoherence. *Sankhyā: The Indian Journal of Statistics, Series A.*, pages 561–587, 2002.

Mark J Schervish, Teddy Seidenfeld, and Joseph B. Kadane. Measures of incoherence: How not to gamble if you must. In *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, page 385. Oxford University Press, USA, 2003.

Miriam Schoenfield. The accuracy and rationality of imprecise credences. *Nous*, page Forthcoming, 2015.

Dana Scott and Peter Krauss. Assigning probabilities to logical formulas. *Aspects of inductive logic*, 43:219–264, 2000.

Teddy Seidenfeld, Mark J Schervish, and Joseph B. Kadane. When fair betting odds are not degrees of belief. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, pages 517–524. JSTOR, 1990.

Teddy Seidenfeld, Mark J Schervish, and Joseph B. Kadane. Forecasting with imprecise probabilities. *International Journal of Approximate Reasoning*, 53 (8):1248–1261, 2012.

Brian Skyrms. Higher order degrees of belief. In *Prospects for Pragmatism*, pages 109–137. Cambridge University Press, 1980.

Julia Staffel. Measuring the overall incoherence of credence functions. *Synthese*, 192(5):1467–1493, 2015.

Johannes Stern. Modality and axiomatic theories of truth I: Friedman-Sheard. *The Review of Symbolic Logic*, 7(02):273–298, 2014a.

Johannes Stern. Modality and axiomatic theories of truth II: Kripke-Feferman. *The Review of Symbolic Logic*, 7(02):299–318, 2014b.

Johannes Stern. Necessities and necessary truths. Proof-theoretically. *Ergo, an Open Access Journal of Philosophy*, 2, 2015a.

BIBLIOGRAPHY

Johannes Stern. *Toward Predicate Approaches to Modality*, volume 44 of *Trends in Logic*. Springer, 2015b. ISBN 978-3-319-22556-2.

William J. Talbott. Two principles of bayesian epistemology. *Philosophical Studies*, 62(2):135–150, 1991.

Alfred Tarski. The concept of truth in formalized languages. *Logic, semantics, metamathematics*, pages 152–278, 1956.

Bas C. Van Fraassen. Belief and the will. *The Journal of Philosophy*, pages 235–256, 1984.

Sean Walsh. Empiricism, probability, and knowledge of arithmetic: A preliminary defense. *Journal of Applied Logic*, 2013.

Brian Weatherson. From classical to intuitionistic probability. *Notre Dame Journal of Formal Logic*, 44(2):111–123, 2003.

J. Robert G. Williams. Generalized probabilism: Dutch books and accuracy domination. *Journal of philosophical logic*, 41(5):811–840, 2012a.

J. Robert G. Williams. Gradational accuracy and nonclassical semantics. *The Review of Symbolic Logic*, 5(04):513–537, 2012b.

J. Robert G. Williams. Probability and non-classical logic. In Christopher Hitchcock and Alan Hájek, editors, *Oxford Handbook of Probability and Philosophy*. Oxford University Press, 2014.

Chunlai Zhou. Belief functions on distributive lattices. *Artificial Intelligence*, 201(0):1 – 31, 2013. ISSN 0004-3702. doi: http://dx.doi.org/10.1016/j.artint.2013.05.003. URL http://www.sciencedirect.com/science/article/pii/S000437021300043X.