A NON-CLASSICAL APPROACH TO MAXIMUM ENTROPY IN UNCERTAIN REASONING

Thomas, Michael R.

2004

MIMS EPrint: **2006.214**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/

And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

A NON-CLASSICAL APPROACH TO MAXIMUM ENTROPY IN UNCERTAIN REASONING

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Science and Engineering

2004

By
Michael R. Thomas
Department of Mathematics

Contents

List of Figures

Abstract

This thesis is concerned with the question “Given a set of knowledge about propositional variables, what is the ‘best’ way to assign probability values to those variables?” I present here an approach to this question based upon a philosophical concept of negation and its role in perception. This concept is discussed in detail before a mathematical analysis of it is presented, in the form of structures in propositional logic which, it is claimed, embody the principles of the underlying philosophy. There follows the definition and mathematical characterisation of an inference process which utilises these logical structures and also adheres closely to the principles of Maximum Entropy. The properties of this inference process are analysed and discussed.

Another inference process is then described based upon a modified version of the philosophical principles defined earlier. A class of graphs is found which are intimately connected with this inference process, and two attempts at characterising this class are presented.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institution of learning.

Copyright

Copyright in text of this thesis rests with the Author. Copies (by any process) either in full, or of extracts, may be made **only** in accordance with instructions given by the Author and lodged in the John Rylands University Library of Manchester. Details may be obtained from the Librarian. This page must form part of any such copies made. Further copies (by any process) of copies made in accordance with such instructions may not be made without the permission (in writing) of the Author.

The ownership of any intellectual property rights which may be described in this thesis is vested in the University of Manchester, subject to any prior agreement to the contrary, and may not be made available for use by third parties without the written permission of the University, which will prescribe the terms and conditions of any such agreement.

Further information on the conditions under which disclosures and exploitation may take place is available from the head of Department of Mathematics.

Acknowledgements

I would like to thank firstly my supervisor Dr. George Wilmers for his stimulating discussion and constant stream of ideas during the period of research leading to this thesis. Without his continued support and guidance this work would never have been finished. I also owe a debt of gratitude to my parents, my family and my many friends who have lent encouragment, inspiration and advice over the past years; in particular Matt, Emily, Chris and Nick. Finally my deepest thanks to Pam for her tolerance, patience and love.

Chapter 1

Introduction

Uncertain reasoning is an area of study that stretches back to the Greek's early attempts to formalise reasoning. In recent years it has become a matter of practicality as interest in *expert systems* and *artificial intelligence* has grown with the explosion of readily available computers. This thesis addresses the question of how to formalise the drawing inferences from uncertain knowledge in a way consistent with a number of philosophical principles concerning uncertainty and perception.

In Chapter 2, I discuss the concept of uncertain reasoning and some of the attempts to formalise it. The principle of maximum entropy is considered in some detail, in preparation for the systems of uncertain reasoning defined later in the thesis. I also consider the nature of negation in the process of perception observable properties in Section 2.2, and propose two logical principles which I believe embody the intuitions discussed in this section. These are the Classification Principle, which states that negative propositions are never observed but only inferred from incompatibilities with positive observed propositions, and the Principle of Conjunctive Closure, which states that consistent conjunctions of observable properties are themselves observable properties.

Chapter 3 formalises the principles discussed in Chapter 2 and presents structures in propositional logic called Positive Frames which capture the philosophical notions discussed earlier. Some results are presented which lend support to the thesis that our intuitions are sound. A 1-1 correspondence between certain types of positive frame and hypergraphs which allows us to use elements of graph theory in our discussions.

In Chapter 4 a probabilistic inference process called CFE is presented which is based upon the principle of maximum entropy in conjunction with the two principles proposed in Chapter 2. The technical definition of the process is accomplished via a restriction of the maximum entropy principle to the logical structures defined in Chapter 3. Two characterisations of the CFE process are presented as evidence for the justifiability of CFE. In Chapter 5 we analyse some of the properties of CFE.

An alternative approach is taken in Chapter 6. Another inference process is defined, this time by considering a strengthening of the Classification Principle and dropping the Principle of Conjunctive Closure, and restricting maximum entropy again, this time to the "normal positive frames" defined by the modified principles. Via the correspondence with graphs defined in Chapter 3, we discover that a certain class of graph is chosen by this inference process. These L-minimal graphs can can be defined as the minimal graphs in a quasi-ordering defined on the class of simple graphs with no isolated vertices. We present two attempts at characterising these graphs — one graph theoretical and one logical. Unfortunately both these attempts are only partial results, with the result that the inference process can only currently be realistically studied for small knowledge sets. A list of the L_1-minimal graphs is found, with proof, and these are used to study the behaviour of the inference process in the one-dimensional case.

Finally, in Chapter 7, we discuss the inference processes and their properties and present our conclusions and recommendations for further research.

Chapter 2

Philosophy & Motivation

We concentrate in this thesis on defining systems of uncertain reasoning and analysing their behaviour mathematically. In Section 2.1 we will briefly discuss uncertain reasoning and some of the attempts at defining and justifying systems of uncertain reasoning. In particular we will be interested in principles of reasoning which define "desirable" properties of such systems. One such principle is the Principle of Maximum Entropy which is fundamental to the systems defined in this thesis. Indeed, our systems will be essentially a combination of this principle combined with a hypothesis about the nature of perception. For this reason we discuss Maximum Entropy in more detail in Section 2.1.1.

In Section 2.2 a hypothesis concerning the nature of perception and the role of negation and negative propositions is advanced. The implications of this hypothesis are discussed and two logical principles are proposed which claim to capture the essence of the hypothesis. These principles are used in Chapter 3 to define types of structure in propositional logic with which we develop systems of uncertain reasoning in Chapter 4 and Chapter 6.

2.1 Uncertain Reasoning

Uncertain reasoning is that area of logic which deals with how we come to conclusions upon uncertain sets of knowledge. If we have uncertain information about a situation how should we choose the "best" representation of it?

Many attempts have been made at formalising and answering these ideas. In this thesis we will be concerned with systems of *probabilistic inference*. These systems attempt make an identification between the uncertain knowledge itself and a set of probability functions on the language in which the knowledge is specified (see [12]. The process of uncertain reasoning is then to pick from this set of probability functions some subset which is considered to be the "best" representation of the knowledge. It should be noted however, that probabilistic inference is not the only way of conducting uncertain reasoning. Non-monotonic reasoning, for example, offers a qualitative method of dealing with inductive inference, while other quantitative efforts at dealing with uncertainty include Dempster-Shafer belief functions ([9]) and possibility theory ([26]), for example. However, we will not go into detail on those theories here.

We have been using "best" in the scare-quoted sense here because there is considerable debate over what constitutes the correct way to pick probability functions. There seems to be no *a priori* correct way to judge what is the best way to pick a set of probability functions to represent our knowledge — indeed there is some debate over the matter. Howson gives in [20] a comprehensive account of the problems involved in inductive inference and the many attempts that have been made to solve them. In [34], Paris describes the problem of probabilistic inference clearly, and goes on to describe a selection of proposed "inference processes."

We remark that in general, the method of attack upon these problems seems to be as follows. Some principle of reasoning is posited and argument are given in justification of it, and then the consequences of assuming that principle are analysed. One such principle is the Principle of Equivalence — this, broadly speaking, says that if two sets of uncertain knowledge are essentially equivalent then any process of uncertain reasoning should arrive at the same conclusions on both sets.

Another important principle is that of maximum entropy, and it is that which we describe in the following section:

2.1.1 The Maximum Entropy Principle

The concept of entropy was introduced in the late 19th century as a concept in statistical physics. In that context, it is regarded as a measure of the disorder of a physical system. The principle of maximum entropy in this situation claims that since there are many more disordered states than ordered ones, we should assume that any system will be in one of the most disordered stated consistent with the observed properties of the system — that is, a state with the maximum entropy.

Shannon introduced the concept of entropy into information theory in the mid 20th century as a measure of the “uncertainty” of a probability function — see [48] for his justification. It proceeds via the assumption of a few simple properties that any measure of uncertainty should satisfy, and from these derives the fact that any such measure must be (a multiple of) the entropy function.

The principle of maximum entropy then states that we should pick, as our representation of uncertain knowledge, that probability function which has greatest uncertainty (i.e., maximum Shannon entropy). The justification given for this is that maximising the entropy is equivalent to maximising the uncertainty in the representation of our uncertain knowledge given by a probability function. That is, we are picking the representation which “makes the least assumptions” beyond that which is contained in the original knowledge set.

Another justification for maximum entropy is given by Jaynes in Chapter 11 of [24], and is also formalised by Paris in [35]. This justification is similar to Boltzmann’s justification of maximum entropy in the physical sense. The idea is that, of all the myriad ways in which a particular set of uncertain knowledge could come about, most of them have high entropy. Jaynes and Paris both show that the vast majority of the situations in which our uncertain knowledge could be generated are “close” (in some sense) to the situation with maximum entropy. Therefore, we should choose the maximum entropy solution as the correct method of representing out uncertain knowledge.

There are many other justifications given for the principle of maximum entropy — for example, Paris also provides a characterisation of this principle as equivalent to the conjunction of several other simpler “common sense” principles in [33]. However, this is not to claim that maximum entropy is without its critics. However we remark here that we feel the principle of maximum entropy to be a sound one on the basis of the justifications mentioned here, and so shall use it as a guiding principle in constructing our own inference processes in Chapters 4 and 6.

2.2 Perception and Negation

The fundamental philosophical motivation for this thesis is a hypothesis about the nature of the process of perception, in particular the role that negation plays in that process. The hypothesis comes from an observation about the way in which we perceive our environment — put roughly, this observation is that we do not seem to actually *observe* or *perceive* negations of attributes. Rather, negations are *deduced* from the totality of our positive perceptions.

The concepts of "positive perception" and "negation of attributes" need some explaining. A positive perception is the reception of information to the effect that some object has a certain property. On the other hand, here the negation of an attribute should be taken to mean that an object does not possess that attribute (or property). Then our hypothesis claims that we do not directly perceive that observable objects do not possess certain properties — instead, we deduce that these properties do not hold from the information we have received about which properties *do* hold for the the given objects[1].

Now, to give a simple example of our hypothesis, consider the statement

$$\text{“The paper on which this page is printed is not red”} \tag{2.1}$$

While the truth of this statement is obvious, in what sense do we infer this from our perceptions? There is no "not red" attribute of the page that we directly perceive. Instead, we perceive a number of things about the colour of the paper — one of which is that it is white. Our contention is that from this perception of the whiteness of the paper, the "not red"-ness is *deduced* via some prior knowledge of incompatibility between redness and whiteness. In other words, we have some prior knowledge about colours of the form

$$\text{“The attributes ‘White’ and ‘Red’ are incompatible”} \tag{2.2}$$

[1] An obvious problem with this hypothesis is the symmetry in most logics between a property and its negation. To address this issue we will make the assumption that the observable properties are "natural" in some sense. Our approach to this problem is covered in the discussion of Natural Kinds in Section 2.2.3

which, when combined with

$$\text{“We perceive that the page is white”} \tag{2.3}$$

allows us to deduce (2.1).

Of course, this is a very simple example — in particular it ignores the many gradations of colour. The incompatibility between red and white seems obvious enough but how would we model, for example, the difference between white and a very light pink? As we make the pink lighter and lighter, at what point does the shade cease to be “red?” These questions are considered in more detail in Pears' *Incompatibilities of Colours* [40], and Gärdenfors' work on conceptual spaces ([13], [14], [15]) attempts to impose more topological structure on attributes.

However, crude as this example may be, it does serve to illustrate our thesis. Despite the difficulties inherent in dealing with “fuzzy” attributes such as colours, certain attributes are mutually incompatible, and we use our knowledge of these incompatibilities to deduce the negation of attributes. Consider as another example an empty coffee mug. I do not directly perceive the emptiness of the mug — I perceive many other things about it though. I can see the bottom of the mug for one thing, and this is incompatible with me having any coffee left to drink. Hence I have deduced the negation of the proposition

$$\text{“There is coffee in my mug”} \tag{2.4}$$

Both of the examples we have considered thus far rely on the visual sense. However, when we refer to “perception” here, we are not simply talking about the act of seeing. Rather, we intend “perception” to be understood as the reception of information about an agent's environment.

Suppose we consider the coffee cup example in a different manner. Suppose I were to dip my fingers in my (still empty) coffee cup: the sensations I would receive from my fingers would be incompatible with the "burning" and "wet" sensations I would receive if there were coffee in the cup. So my perception of my environment would be incompatible with the proposition (2.4), which we would deduce to be untrue. Hence we are indeed talking about perception rather than just "seeing."

Note that in the previous example, it may be argued that we have already denied our thesis. Surely I am claiming that I have *perceived* a lack of sensation from my fingers, and thus I am in fact directly perceiving the negation of an attribute? On the contrary though, I am claiming no such thing. In fact, I am claiming that there are (at least) two distinct positive sensations I may perceive from my fingers. The most obvious "direct sensation" is that of being dipped in scalding hot coffee — no doubt this is something I would directly perceive from my fingers. The second, and more subtle sensation, is that of "normality", where my fingers are at a comfortable temperature and are dry. While this sensation is unarguably less urgent than the first perception, I claim it is something that I directly perceive nonetheless — it is *not* a lack of perception. In this case, a lack of perception would correspond to the loss of *all feeling* in my fingers, a very different state from that of the (almost subconscious) sensation of one's fingers feeling "normal."

It is worth expanding upon this point. We do so with another example that shows also that we are not restricting ourselves solely to human, or even animal, perception. In this example we consider a computer receiving information from some device as a string of 0s and 1s — perhaps a keyboard on which someone is typing, or maybe a modem connected to the Internet. It is clear that when the computer receives a 0, it has not perceived "the absence of a 1" — on the contrary the computer has directly perceived the presence of a certain voltage level in some circuit which it interprets as a 0. The absence of any information being passed to the computer would not mean that the computer would "perceive" a 0. When formulated this way, we can clearly see that there is a difference between a lack of perception and "default" perceptions.

2.2.1 Perspectives on the Philosophy of Negation

This section relies heavily on the excellent survey of the many varying views of negation proposed through history by Cleave in [5].

The example of the computer discussed at the end of the previous section suggests a 3-valued logic may be appropriate for modelling our putative principle. While we will in fact use classical propositional logic for the main body of this thesis, a search of the literature on many-valued logics reveals that N.A. Vasil'ev proposed a similar principle in 1913 ([51]). Kline's investigation [25] of this logician's work includes Vasil'ev's claim that

> "...negative predicates are not primitive but are inferred from positive predicates. Negative propositions concerning perceptions are inferred from propositions about incompatible properties: a denial that an object has a property P is founded on the the presence of a witness having a property N which excludes P."[2]

This claim is very similar to that which we make at the beginning of this section. In fact, the concept of the existence of a witness possessing an incompatible property will be crucial to our discussion. While Vasil'ev's work concentrated on many-valued logics however, we will use these intuitions to develop systems of uncertain reasoning.

[2] Quoted from p. 112 of [5].

Cleave's survey [5], from which the above quotation is taken, also discusses some of Wittgenstein's views on negation. Negation, Wittgenstein believed, is "not something in the world," although it is "constitutive with reality in some sense," a view which is consistent with the argument we are putting forward here. For we claim that a negative fact cannot be perceived as part of the observable world, even though it is true. However, the fact that we may (possibly) infer that fact from the properties of the world that we have observed indicates that it does have some connection with reality.

Cleave cites Wittgenstein's *Tractatus* ([55]) in a context which is important to our argument:

> (2.04) The totality of existent atomic facts is the world.
>
> (2.05) The totality of existent atomic facts also determines which atomic facts do not exist.
>
> (2.06) The existence and non-existence of atomic facts is the reality.
>
> (The existence of atomic facts we also call a positive fact, their non-existence a negative fact.)

This sits well with our hypothesis of positive perceptions. Our positive perceptions correspond to Wittgenstein's existent atomic facts, and in turn they determine the non-existent atomic facts — statements about properties which do not hold in our framework.

However, Priest challenges our hypothesis directly in [42]. Addressing Vasil'ev's contention that we cannot perceive something that is not the case, Priest disagrees:

> When we perceive, we can see that something *is* the case. Can we also perceive that something is *not* the case? Some have thought

> not. We can only perceive that something *is* the case. For example, we cannot see that something is not green. We can only see that it is red. Any judgement to the effect that it is not green has to be added to what we see by inference. This, as we now see, is false. I can see directly that something is not green. Or consider another example: you enter a room; the whole room is visible from where you stand; there is no one there. You can see that Pierre is not in the room. No Pierre-shaped objects meet the eye[3].

I disagree with this criticism on two counts. Firstly, it is not clear what it means to "see *directly* that something is not green." Seeing that something *is* green seems to be a reasonable candidate for something we can perceive directly — we could define it as being the event of light of a certain minimum intensity, whose wavelength falls within a certain range, falling upon one's retina.

However, to see directly that something is not green does not seem to admit of such a convenient definition. To be sure, we could define the act of seeing directly that something is not green as being the event whereby light of a certain minimum intensity, whose wavelength is *outside* a certain range, falls upon one's retina. The problem with this is that even this apparently simple definition relies upon there being an *a priori* incompatibility between those wavelengths called green and those called not-green. We see that the problem of direct perception here has simply been removed to a different level of abstraction.

To answer Priest's (and Sartre's) second example — that of the absence of Pierre — we turn to Wittgenstein's conception of the totality of existent atomic facts. Priest claims that we can directly perceive the absence of Pierre because "no Pierre-shaped object meets the eye." But by what means do we know this?

[3] Priest notes that this second example comes from Sartre [47]

It is simply by the fact that we have directly perceived a *totality* of positive stimuli about the room, the sum total of which is incompatible with any Pierre-shaped object being observed. So we have in fact inferred Pierre's absence from that which we have observed — we have not perceived it directly at all. In Wittgenstein's terminology, according to (2.05) above, it is the *totality* of existent atomic facts (i.e. the observed world) that determines the non-existent atomic facts — in particular, that Pierre is absent.

Our critic may not accept this argument though. A possible counter-argument could be that we might mentally list all the objects that we have observed, and note that no Pierre-shaped objects are in that list. Again though, this is a process of inferring knowledge from what has been perceived — not direct perception itself. Another answer to this argument turns it on its head: we need only to observe that no "absence-of-Pierre"-shaped object meets the eye either. So how are we justified in concluding the absence or otherwise of Pierre? It must be because we have some *a priori* incompatibility between the presence of Pierre and the list of observed shapes, which we use to infer his absence.

This last argument raises an important point about perception — is it in fact a consistent process? In other words, can we observe contradictions? This question, and some of the debate surrounding it, is discussed in the next section.

To further clarify matters, we note a distinction made by Beall and Colyvan in [2] on the difference between weakly and strongly observable properties. They make the definition that a state of affairs σ is called a) *weakly observable* iff it can be observed that σ holds, and b) *strongly observable* iff that state of affairs itself can be observed. In this thesis, we note that we consider the act of perception to correspond to observing *strongly observable* properties. That is, that we perceive only that which we can directly observe — any further information is inferred from these perceptions. Thus, observing that an object is green is a case of strong observability, whereas deducing from this fact that it is not-red is a case of weak observability: we have observed that the "not-red" state of affairs holds, but we have not directly perceived the "not-red" property.

Finally, Priest raises another important point — the question of distinguishing between attributes and their negation, and which of these we can reasonably call positive perceptions. This question is addressed in Section 2.2.3

2.2.2 Observable Contradictions

The concept of negation in logic has a long and controversial philosophical history, one which is tied intimately to the concept of contradiction and stretches back to Greek philosophers such as Zeno, Plato and Aristotle. Some recent discussion on "observable contradictions" and the question of whether the observable world is consistent would seem to be relevant this exposition. We are basing our argument upon the assumption that we can only acquire "positive perceptions," and that it is not possible to perceive a negation.

The question of whether or not there are true contradictions has attracted some debate recently. Dialetheists, such as Priest, hold that there are such things as true contradictions — for a good survey of the arguments proposed, see [43]. We are not concerned overly with the question of whether or not there are such things here, but an important point does get raised in this argument. If we accept that there are true contradictions, it is natural to ask what they would look like, and whether or not we do see any. This is certainly important to the argument presented here — if it were possible to observe a contradiction, this would count as perception of a negation. Such a perception would of course discredit our hypothesis.

Priest claims in [42] that whether or not there are true contradictions, they are not observable. That is, the observable world is consistent. The argument runs along the lines that if there are observable contradictions, we would observe them, and since we do not observe any contradictions then the observable world must be consistent.

As Beall and Colyvan point out in [1] and [2], this argument relies on the assumptions that observable contradictions would be observed, that we would recognise a contradiction if we saw one, and that we do not in fact observe any contradictions. They criticise these assumptions and give arguments to claim that we do in fact observe contradictions. In fact, they claim that contradictions are *strongly observable*, in the sense described in the last section, and describe the first steps toward a para-consistent theory of vagueness, as posited by Hyde ([21]), to outline their argument.

However, I would argue that, if there are true contradictions, then they are only *weakly observable* — that is, that we can infer that a contradiction is true from what we directly perceive, but that we cannot directly perceive a contradiction (true or not). The reasoning behind this argument is that to observe a contradiction, we must directly perceive that something is the case, and that it is also not the case. Leaving aside the question of whether or not it is possible to directly perceive something which is not the case (and of course we argue that it is not), if we were to directly perceive that some property holds of an object, and also that it does not hold of that same object, this is not the same as directly perceiving the contradiction that it does and does not hold of that object. Rather the observation that a contradiction is true of that object is an inference from the two direct perceptions and their incompatibility. In this sense, a contradiction is only weakly observable.

In other words, if we were to perceive that an object is green, and also to perceive that it is not-green, then we could infer that it is green and not-green —

a contradiction. But this is only a contradiction by the incompatibility of green and not-green.

In summary then, I would hold that a weaker version of Priest's hypothesis that the observable world is consistent is true. Drawing on Beall and Colyvan's arguments, I suggest that in fact the strongly observable world is consistent, but the weakly observable world may not be.

2.2.3 Observable and Inferred Properties

We mentioned at the beginning of this section (see footnote on page 16)and in subsection 2.2.1 that there is a problem with the concept of perception as we have thus far outlined it, one that is due to the inherent symmetry in naming properties. Priest for example claims in [41] that

> . . . the very distinction between seeing what is the case and what is not the case is a false one. Some seeings are both. With respect to physical objects, to be transparent is not to be opaque, and vice versa. But you can see that something is transparent and you can see that something is opaque.

This is indeed a problem for our embryonic systems. It is clear from our earlier discussions of the nature of perception that we require some distinction to be made amongst properties. Some properties are intuitively more "natural" than others in this context.

One concept seems to be quite obviously applicable to this problem — the idea of "natural kinds." This is a concept which dates back to Aristotle (see [16]), and which essentially refers to classes of objects which are "natural" in some way. They are classes which share some underlying commonality — according to Wilkerson ([53]), "An object is a member of a natural kind in virtue of having a real essence: a set of properties necessary and sufficient for membership of that kind."

There is some difficulty in the definition of natural kinds though. Quine, for example, explains in [46] that many notions in philosophy can be definable in terms of natural kinds, but that any general definition of 'natural kind' is not possible. Theories of natural kinds have also been developed by Kripke ([27], [28]) and Putnam ([44]), but theories of natural kinds often come under attack for (Aristotelian) essentialism - "the doctrine that some of the attributes of a thing . . . may be essential to the thing, and others accidental." ([45]

For our purposes, we do not concern ourselves overly with the exact definition of natural kinds here. It is sufficient for us to note that natural kinds are classes of object which share some common necessary and sufficient properties (beyond that of belonging to the same class). We simply note that some classes are "natural" and examine the implications for our theory. To take one of Hardegree's examples in [18], of the two classes

1. the class of all humans;

2. the class of consisting of Mozart, the planet Jupiter, and the number 41

the first class seems, intuitively, more natural than the second.

The logic that Hardegree proposes for natural kinds in [18] is based on the supposition that there are "traits," or properties that can hold for objects. A natural class is a class for which there is a set of traits such that all and only members that class share that particular set of traits. Hardegree goes on to develop a logic of natural classes and traits which, under a particular Galois connection assigning sets of traits to the classes that hold them, form a lattice structure. A natural kind is then an ordered pair consisting of a set of individuals and a set of traits such that the traits are those corresponding to the class of individuals, and the individuals form that natural class determined by the set of traits.

It seems obvious to me that strongly observable properties must be instances of Hardegree's traits. As Cocchiarella points out in [6], "natural kinds are *material* and not *logical* essences." That is, traits are properties of the objects in the world, and this must include strongly observable properties.

We remark then that we assume for the purposes of this thesis that our properties are "natural" in the sense discussed here.

2.2.4 Some logical principles

We describe in this section some more formal logical principles based on the discussion thus far. Firstly we formalise our fundamental insight as the following principle, phrased around Kline's interpretation of Vasil'ev's work in [25]:

Definition 2.1 (The Classification Principle).
The assertion that an object does not have a property P is only effected by the assertion that it has a property N, or conjunction of such properties, which is incompatible with P.

This principle will be important in defining the structures of propositional logic which we present in the next chapter and which we propose capture the notions discussed here. Notice that we do not insist that N is a directly perceived property. N itself may have been inferred from some directly perceived information, or from other inferred properties.

The second part of this definition is important too. For it may be the case that some combination of properties is incompatible with P, but that that particular conjunction has no name in the language of discourse.

Note that we insist upon conjunction here — we do not allow disjunction. This arises from consideration of the logic of natural kinds as discussed in the previous section. Following a similar argument to that of Hardegree in his logic of natural kinds, it seems clear that the conjunction of two or more (strongly) observable properties would itself be (strongly) observable — that is, if we could directly perceive α and β then we could also directly perceive $\alpha\&\beta$. However, the disjunction of two such properties does not seem to determine a strongly observable or natural property at all. Consider for example, the two properties "green" and "Pierre." In our previous discussions in this chapter we have argued that these are directly perceivable properties. However, it seems clear that there is no sensible conception of "green or Pierre" as a directly perceivable property, especially in light of our consideration of observable properties as being "natural" in some sense.

Our second principle arises from consideration of the second part of the Classification Principle, and also from the logic of observable properties. If, as we say, some combination of observable properties can determine the denial of some other property, then surely that conjunction, being observable, should have a name:

Definition 2.2 (The Principle of Conjunctive Closure).

Any consistent conjunction of observable properties is an observable property.

Chapter 3

Positive Frames

In this chapter we define some logical structures which we claim capture the intuitions discussed in the previous chapter. We work here with finite propositional languages, which although lacking in expressive power, have the advantage of being simple to work with. In some of the technical discussions of Chapters 4–6, this will prove to be very important in keeping our discourse coherent. Explanations of propositional logic can be found in most textbooks on logic, such as [10] for example.

Of course, we can always consider the propositions to be unary predicates to be applied to a single individual at a time. This sits well with the discussion in the previous chapter, where we considered observable properties of objects in the world. Throughout this chapter we consider the propositions of our language to be observable properties in the sense discussed in Chapter 2.

We begin with the definitions that we will require to build the logical structures encompassing the principles in Section 2.2.4.

3.1 Basic Concepts

We begin with the idea that the process of perception corresponds to the reception of some positive information. We start with a weaker principle than those discussed in Chapter 2.2.4, and consider a perception to be some statement about only positive properties which hold.

Definition 3.1 (Positive Sentences).

In a language L for propositional logic a **positive sentence** is a sentence of L in which no negation or implication signs occur.

More precisely, in a language L, the set of positive sentences $PL \subseteq SL$ is defined inductively by:

1. $PL^0 := L\,,$
2. $PL^{k+1} := \{\theta \,\Box\, \phi \;\mid \theta \in PL^i,\ \phi \in PL^j \text{ for some } i, j \leq k,\ \Box \in \{\wedge, \vee\}\,\}$

and finally set

$$PL := \bigcup_{k=0}^{\infty} PL^k\,.$$

We now define a framework in which everything can be related to the positive perceptions described by the previous definition.

Definition 3.2 (General Positive Frames).

A **general positive frame** on a propositional language L is a theory T of SL relative to which all sentences of L are equivalent to a positive sentence. That is, T is such that for all $\theta \in SL$ there is some $\theta^* \in PL$ for which

$$T \models \theta \leftrightarrow \theta^*\,.$$

This corresponds to the notion that we must infer everything we know from only our positive perceptions. The next definition, that of an atom of a positive frame, will be very important in the technical discussions to follow later in this thesis.

Definition 3.3 (Atoms of positive frames).

Take L to be a finite propositional language. We follow the usual convention[1] of calling a maximal conjunctions of literals of L an **atom** of L. That is, an atom of L is any conjunction of the form

$$\bigwedge_{p \in L} p^{\epsilon_p}$$

where $\epsilon_p \in 0, 1$ and $p^1 = p$ while $p^0 = \neg p$.

Notice that if $L = L_n$ then there are 2^n atoms of L_n. In this thesis we will denote by $\alpha_1, \alpha_2, \ldots, \alpha_{2^n}$ the atoms of L_n unless otherwise stated.

We denote the set of all atoms of L by $\mathrm{At}(T)$. For any general positive frame T on L, denote by $\mathrm{At}(T)$ the set of atoms of L which satisfy T.

The General Positive Frames described above give us a framework in which we can discuss positive perceptions. However, we would like to be able to formulate our logical principles of Section 2.2.4 in such a way as to allow us to investigate their formal consequences. We begin with the Classification Principle.

Definition 3.4 (The Classification Principle).

Within a propositional language L we can specify the Classification Principle of Definition 2.1 in the following way. For some $p \in L$, call any sentence of the

[1] See, for example, [34] page 13

following type a **classification sentence for p**

$$\neg p \leftrightarrow \bigvee_{i=1}^{k} \bigwedge W_i$$

when $k \geq 1$ and the sets $W_1, W_2, \ldots, W_k$ are s.t. $\emptyset \neq W_i \subseteq L \setminus \{p\}$.

Call the sets $W_1, \ldots, W_k$ **witness sets for ¬p**. If there is some W_i s.t. $W_i = \{q\}$ then call q **a witness for ¬p**.

Now we consider the arguments about totality of perception to give the following definitions

Definition 3.5 (Positive Frames).

A **positive frame** is a theory T of SL which has a classification sentence for every $p \in L$.

That is, if $L = \{p_1, p_2, \ldots, p_n\}$, say, then there are $k_1, \ldots, k_n \geq 0$ and for every $1 \leq i \leq n$ there are sets $W_{i,1}, W_{i,2}, \ldots, W_{i,k_i}$ such that

$$T \models \neg p_i \leftrightarrow \bigvee_{j=1}^{k_i} \bigwedge W_{i,j}$$

and $\emptyset \neq W_{i,j} \subseteq L \setminus \{p_i\}$.

The next definition imposes some more conditions on these structures.

Definition 3.6 (Reflexive Classification Sentences).

A set of classification sentences on L_n is said to be **reflexive** if there is exactly one classification sentence for each $p \in L_n$ and the following conditions hold:

1. For every $i, i' = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, k_i$

$$p_{i'} \in W_{i,j} \;\Rightarrow\; \exists 1 \leq j' \leq k_{i'} \text{ s.t. } W_{i',j'} = (W_{i,j} \setminus \{p_{i'}\}) \cup \{p_i\} \;;$$

2. For every $i = 1, 2, \ldots, n$ and $j, l = 1, 2, \ldots, k_i$ if $j \neq l$ then $W_{i,j} \not\subseteq W_{i,l}$.

These seem reasonable conditions to us. Surely if we believe that P is incompatible with N, then we must also take N incompatible with P. The second condition simply says that if we already consider the conjunction of a set W to be incompatible with P, then there is no reason to include any supersets in the definition of P's incompatibilities.

Corollary 3.1 *Notice that from the second condition in the above definition a stronger condition actually follows: Namely that for every $i, j \in 1, 2, \ldots, n$, $r \in 1, 2, \ldots, k_i$ and $s \in 1, 2, \ldots, k_j$ we have $W_{i,r} \cup \{p_i\} \not\subset W_{j,s} \cup \{p_j\}$.*

Proof. Suppose $W_{i,r} \cup \{p_i\} \subset W_{j,s} \cup \{p_j\}$. Then $p_j \in W_{i,r}$ and so, since this is a reflexive set of classification sentences there must be some $t \in 1, 2, \ldots, k_j$ such that $W_{j,t} = (W_{i,r} \setminus \{p_j\}) \cup \{p_i\}$. Hence

$$\begin{aligned} & W_{i,r} = (W_{j,t} \setminus \{p_i\}) \cup \{p_j\} \\ \Rightarrow\quad & W_{j,t} \cup \{p_i, p_j\} \subset W_{j,s} \cup \{p_j\} \\ \Rightarrow\quad & W_{j,t} = W_{j,t} \cup \{p_i\} \subset W_{k,l} \end{aligned}$$

which is a contradiction to the 2nd condition of Definition 3.6. □

For technical reasons we will make a connection between the sets of reflexive classification sentences and the theories they determine:

Definition 3.7 (Normal Positive Frames).

A **normal positive frame** on a propositional language L is a theory of L which is generated by a reflexive set of classification sentences for L.

There is another property of positive frames which seems relevant to us: that of contingency. In the discussion in chapter 2 we outlined our theory of observable properties. There is an argument to be made that properties that are either ever-present or never observed are not observable properties at all. It seems obvious that a property that is never observed is not an observable property, but what of ever-present properties? I would contend that a property which is always the case cannot be proper subject of a perception: a logical necessity is something that is deduced. It is unclear what it would mean to *directly perceive* a logical truth. For this reason we make the following definition:

Definition 3.8 (Contingent Frames).

A general positive frame T on a propositional language L is called **contingent** if for every for every $p \in L$ we have both $T \not\models p$ and $T \not\models \neg p$.

3.1.1 Relationships between types of Positive Frame

In this section we explore some of the relationships between the different types of structures defined above, and study some of their properties.

To begin with, it turns out that even in their most general form, positive frames allow us to consider statements as being equivalent to a disjunction of some conjunctions of positive perceptions. In other words, we have a sort of "positive normal form" for sentences in General Positive Frames:

Lemma 3.2 *If θ is a positive sentence of L then there are $k \geq 1$ distinct sets $\emptyset \neq S_1, \ldots, S_k \subseteq L$ s.t.*

$$\models \theta \leftrightarrow \bigvee_{i=1}^{k} \bigwedge S_i \,.$$

Proof. We prove the claim by induction on the complexity of θ. If $\theta \in PL^0$ then $\theta = p$ for some $p \in L$, and so the claim is true.

Now suppose that the claim is true for all $\theta \in \bigcup_{i=0}^{r} PL^r$ for some $r \geq 0$, and take $\theta \in PL^{r+1} \setminus PL^r$.

By construction of PL, either $\theta = \phi \vee \psi$ or $\theta = \phi \wedge \psi$ for some $\phi, \psi \in \bigcup\limits_{j=0}^{r} PL^j$. By hypothesis we have $k_1, k_2 \geq 1$ non-empty sets $S_1, \ldots, S_{k_1}$ and $T_1, \ldots, T_{k_2}$ s.t.

$$\models \phi \leftrightarrow \bigvee_{i=1}^{k_1} \bigwedge S_i$$

and

$$\models \psi \leftrightarrow \bigvee_{i=1}^{k_2} \bigwedge T_i$$

where $\mathcal{S} = \{S_1, \ldots, S_{k_1}\}$ and $\mathcal{T} = \{T_1, \ldots, T_{k_2}\}$ are collections of pairwise distinct sets.

Alternatively, if $\theta = \phi \vee \psi$ then

$$\begin{aligned} & \models \theta \leftrightarrow (\bigvee_{i=1}^{k_1} \bigwedge S_i) \vee (\bigvee_{i=1}^{k_2} \bigwedge T_i) \\ \Rightarrow \quad & \models \theta \leftrightarrow \bigvee_{X \in \mathcal{S} \cup \mathcal{T}} \bigwedge X \end{aligned}$$

and so the claim is true for $\theta = \phi \vee \psi$.

If $\theta = \phi \wedge \psi$ then

$$\begin{aligned} & \models \theta \leftrightarrow (\bigvee_{i=1}^{k_1} \bigwedge S_i) \wedge (\bigvee_{i=1}^{k_2} \bigwedge T_i) \\ \Rightarrow \quad & \models \theta \leftrightarrow \bigvee_{i=1}^{k_1} \bigvee_{j=1}^{k_2} (\bigwedge S_i \wedge \bigwedge T_j) \\ \Rightarrow \quad & \models \theta \leftrightarrow \bigvee_{i=1}^{k_1 k_2} \bigwedge R_i \end{aligned}$$

where $R_{(i-1)k_2+j} = S_i \cup T_j$. Clearly we can eliminate any repetitions from this list, and so the claim is true for $\theta = \phi \wedge \psi$.

Hence the claim is true for any $\theta \in PL^{r+1} \setminus PL^r$ and so by induction the Lemma is proved for all $\theta \in PL$. □

Now, the argument about contingency of frames turns out to be important. If a general positive frame is contingent then we can express it as a positive frame — in other words, the Classification Principle holds for contingent general positive frames.

Proposition 3.3 *If T is a contingent general positive frame on L_n, then T is a positive frame on L_n.*

Proof. Since T is a general positive frame then for each $1 \leq i \leq n$ there is $\theta_i \in PL_n$ such that

$$T \models \neg p_i \leftrightarrow \theta_i \,.$$

Then by Lemma 3.2 for each $i = 1, 2, \ldots, n$ there are $k_i \geq 1$ distinct non-empty sets $W_{i,1}, W_{i,2}, \ldots, W_{i,k_i}$ such that

$$\models \theta_i \leftrightarrow \bigvee_{j=1}^{k_i} \bigwedge W_{i,j}$$

Hence we have

$$T \models \neg p_i \leftrightarrow \bigvee_{j=1}^{k_i} \bigwedge W_{i,j} \,.$$

Now it remains only to show that $p_i \notin W_{i,j}$. So for each $i = 1, 2, \ldots, n$ let

$$l_i = |\, \{ W_{i,j} \ \mid p_i \notin W_{i,j} \} \,| \,.$$

Notice that since T is contingent $l_i > 0$. For if not, then let $V_{i,j} = W_{i,j} \setminus \{p_i\}$

for $j = 1, 2, \ldots, k_i$. Then we have

$$\begin{aligned} & T \models \neg p_i \leftrightarrow \bigvee_{j=1}^{k_i} [p_i \wedge \bigwedge V_{i,j}] \\ \Rightarrow \quad & T \models \neg p_i \leftrightarrow p_i \wedge [\bigvee_{j=1}^{k_i} \bigwedge V_{i,j}] \\ \Rightarrow \quad & T \models p_i \end{aligned}$$

which contradicts T being contingent.

So we have $1 \leq l_i \leq k_i$. Relabel the $W_{i,j}$'s so that $p_i \in W_{i,j} \Leftrightarrow j > l_i$ and for $j > l_i$ let $V_{i,j}$ be as above. Then

$$\begin{aligned} & T \models \neg p_i \leftrightarrow \left(\bigvee_{j=1}^{l_i} \bigwedge W_{i,j} \right) \vee \left(p_i \wedge \bigvee_{j=l_i+1}^{k_i} \bigwedge V_{i,j} \right) \\ \Rightarrow \qquad & T \models \neg p_i \rightarrow \bigvee_{j=1}^{l_i} \bigwedge W_{i,j} \, . \end{aligned}$$

Now note that

$$\begin{aligned} & T \models \left(\bigvee_{j=1}^{k_i} \bigwedge W_{i,j} \right) \rightarrow \neg p_i \\ \Rightarrow \qquad & T \models \left(\bigvee_{j=1}^{l_i} \bigwedge W_{i,j} \right) \rightarrow \neg p_i \end{aligned}$$

and so

$$T \models \neg p_i \leftrightarrow \bigvee_{j=1}^{l_i} \bigwedge W_{i,j} \, .$$

Hence T is indeed a positive frame on L_n. □

3.2 Hypergraphs

We will find there is a close relationship between positive frames and hypergraphs, which allows us to study more properties of positive frames. It will also turn out to be a very important relationship in the definition of inference processes in Chapters 4 and 6. We take this definition from Berge [3],(pp.3):

Definition 3.9 (Hypergraph).

Let $X = \{x_1, x_2, \dots, x_n\}$ be a finite set. A **hypergraph** on X is a family $H = (E_1, E_2, \dots, E_m)$ of subsets of X such that:

1. $E_i \neq \emptyset \quad (i = 1, 2, \dots, m)\,;$

2. $\cup_{i=1}^{m} E_i = X\,.$

   A **simple hypergraph** (or "Sperner family") is a hypergraph $H = (E_1, E_2, \dots, E_m)$ such that

3. $E_i \subseteq E_j \;\Rightarrow\; i = j\,.$

The elements $x_1, x_2, \dots, x_n$ of X are called *vertices*, and the sets $E_1, E_2, \dots, E_m$ are the *edges* of the hypergraph. A simple graph is a simple hypergraph each of whose edges has cardinality 2; we shall not consider isolated points of a graph to be vertices.

A hypergraph H may be drawn as a set of points representing the vertices. The edge E_j is represented by a continuous curve joining the two elements if $|E_j| = 2$, by a loop if $|E_j| = 1$, and by a simple closed curve enclosing the elements if $|E_j| > 3$.

The **anti-rank** of a hypergraph is the minimum cardinality of its edges ([3], pp. 4).

For the purposes of this discourse we will be concerned with simple hypergraphs of anti-rank 2. The following concept from graph theory will be of central importance:

Definition 3.10 (Maximal Independent Set).
Given a hypergraph $H = (E_1, E_2, \ldots, E_m)$ on a set X, $A \subseteq X$ is called an **independent set** of H if there is no $E_i \in H$ such that $E_i \subseteq A$.

A is called a **maximal independent set** of H if there is no independent set B of H such that $A \subset B$.

Theorem 3.4 *There is a 1-1 correspondence between reflexive sets of classification sentences and simple hypergraphs of anti-rank 2.*

Proof. We give a construction for a mapping from the set of reflexive sets of classification to simple hypergraphs of anti-rank 2, and a construction for its inverse:

1. Let X be a reflexive set of classification sentences on L_n. For each $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k_i$ define $E_{i,j} = W_{i,j} \cup \{p_i\}$, and let

$$H^* = \{V_{i,j} \mid i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, k_i\} .$$

Now consider $V_{i,j} \in H^*$ of cardinality k. Since X is a reflexive set of classification sentences H^* will contain exactly $k-1$ other sets $V_{i',j'} = V_{i,j}$. Let H be H^* with all such repetitions deleted.

Then H is a hypergraph on L_n: indeed, it is obvious that $V_{i,j} \neq \emptyset$. Secondly, each $p_i \in L_n$ appears in $\sum_{j=1}^{k_i} |V_{i,j}|$ sets of H^*, of which $\sum_{j=1}^{k_i} |V_{i,j}| - 1$ are deleted to form H. Hence every $p_i \in L_n$ appears in some set of H and so

$$\bigcup_{V\in H} V = L_n \, .$$

Suppose now that we have $E, F \in H$ such that $E \subset F$; then there must be some $p_i \in E \cap F$. By construction of H we must have some distinct j, j' such that $E = W_{i,j} \cup \{p_i\}$ and $F = W_{i,j'} \cup \{p_i\}$. This gives $W_{i,j} \subset W_{i,j'}$, which is a contradiction to the second condition for the reflexivity of the set X. Hence H is a simple hypergraph, and is also trivially of anti-rank 2 since $W_{i,j} \neq \emptyset \Rightarrow |V_{i,j}| \geq 2|$.

2. Take H to be a simple hypergraph of anti-rank 2 on L_n, and for each $i = 1, 2, \dots, n$ let $k_i = |\{E \in H \ | \ p_i \in E\}|$. Set $D_{i,1}, D_{i,2}, \dots, D_{i,k_i}$ to be the k_i distinct edges of H which contain p_i. We now define a set of classification sentences for L_n by setting $W_{i,j} = D_{i,j} \setminus \{p_i\}$ and putting

$$X = \left\{ \neg p_i \leftrightarrow \bigvee_{j=1}^{k_i} \bigwedge W_{i,j} \ \mid 1 \leq i \leq n \right\}$$

Note first that X is indeed a set of genuine classification sentences since each p_i must be in some edge of H and so $k_i \geq 1$. Also H being of anti-rank 2 implies that $|D_{i,j}| \geq 2$, and so $W_{i,j} \neq \emptyset$.

We check now the two conditions for reflexivity given in Definition 3.6:

(a) Suppose that for some $i' \neq i$ we have $p_{i'} \in W_{i,j}$ for some j. Then by construction of X there is an edge E of H such that $E =$

$W_{i,j} \cup \{p_i\}$. Since $p_{i'} \in E$ then there is some $D_{i',j'} = E$. Then

$$\begin{aligned} W_{i',j'} &= S_{i',j'} \setminus \{p_{i'}\} \\ &= D_{i,j} \setminus \{p_{i'}\} \\ &= (W_{i,j} \cup \{p_i\}) \setminus \{p_{i'}\} \\ &= (W_{i,j} \setminus \{p_{i'}\}) \cup \{p_i\} \quad \text{as required.} \end{aligned}$$

(b) Suppose for some i there are j, l such that $W_{i,j} \subseteq W_{i,l}$. Then there are edges $E_1 = W_{i,j} \cup \{p_i\}$ and $E_2 = W_{i,l} \cup \{p_i\}$ of H for which $E_1 \subseteq E_2$. But H is simple and so we must have $E_1 = E_2$. Hence $W_{i,j} = W_{i,l}$ and so $j = l$.

It is easy to check that each of the constructions given above defines an injective mapping and that the mappings so defined are the inverse of each other. Hence the 1-1 correspondence is established. □

This theorem shows us that there is a very nice graphical representation of reflexive sets of classification sentences. However, we are more interested in the consequences of such reflexive sets, so we would prefer to have a graphical representation of normal positive frames. We will now construct such a representation. Firstly, we construct a link between maximal independent sets of a hypergraph and the valuations of its corresponding reflexive set of classification sentences.

Proposition 3.5 *Let X be a reflexive set of classification sentences for L_n, let H be the hypergraph generated by X, and let A be a subset of L_n. Then the valuation v defined on L_n by setting*

$$v(p_i) = 1 \Leftrightarrow p_i \in A$$

is consistent with X iff A is a maximal independent set of H.

Proof. $\Rightarrow$ Let v be a valuation on L_n consistent with X. Suppose the set $A \subseteq L_n$ defined by v is not independent. Then there is some edge E of H such that $A \supseteq E$. By construction of H then there is some $p \in L_n$ and some witness set W for $\neg p$ such that $W \cup \{p\} \subseteq A$. Hence $v(p) = v(\bigwedge W) = 1$. But since W is a witness set for $\neg p$, this contradicts v being consistent with X, and so A is indeed independent w.r.t. H.

Suppose then that A is independent w.r.t H, but not maximal such. Then there is some $p \in L_n \setminus A$ such that $A \cup \{p\}$ is independent. But $p \notin A \Rightarrow v(p) = 0$. Since v is consistent with X then this gives us $v(\bigwedge W) = 1$ for some witness set W for $\neg p$.

By construction of H though, there exists some edge E of H such that $E = W \cup \{p\}$. Then $E \subseteq A \cup \{p\}$, which contradicts the independence of $A \cup \{p\}$, and so A is indeed maximally independent.

$\Leftarrow$ Let A be a maximal independent set of H. Consider $p_i \in L_n$. If $p_i \in A$ then $v(p_i) = 1$, and so for v to be consistent with X we must have $v(\bigvee_{j=1}^{k_i} \bigwedge W_{i,j} = 0$.

That is, for every $j = 1, 2, \ldots, k_i$ we must have $v(\bigwedge W_{i,j}) = 0$. Suppose not: then there is some $W_{i,j}$ such that $v(p) = 1$ for all $p \in W_{i,j}$. Then by construction of H there is an edge $E = \{p_i\} \cup W_{i,j}$ of H such that $E \subseteq A$, contradicting the independence of A. So v is indeed consistent with the classification sentence for p_i.

Now suppose that $v(p_i) = 0$. To ensure v's consistency we must then have $v \bigvee_{j=1}^{k_i} \bigwedge W_{i,j}) = 1$. Suppose not: then for every $j = 1, 2, \ldots, k_i$ we have at least one $p \in W_{i,j}$ such that $v(p) = 0$. That is, for every witness set W for $\neg p_i$ there is at least one $p \in W$ for which $p \notin A$. Then $A \cup \{p_i\}$ is independent w.r.t H: since for every edge E of H which contains E there is at least one $p \in E$ such that $p \notin A \cup \{p_i\}$, and any edge of E which does not contain p_i is not a subset of $A \cup \{p_i\}$ by the independence of A. This contradicts the maximality of A, and so v is again consistent with the classification sentence for p_i.

Hence for all $\theta \in X$, $v(\theta) = 1$.

□

Corollary 3.6 *Let X be a reflexive set of classification sentences on L_n. Then for any $A \subseteq L_n$, $X \models \neg \bigwedge A$ iff there is some $p \in L_n$ and witness set W for $\neg p$ such that $A \supseteq W \cup \{p\}$.*

Further, if A is such that for any $B \subset A$ we have $X \not\models \neg \bigwedge B$ then in fact $A = W \cup \{p\}$.

Proof. $\Rightarrow$ Let H be the hypergraph generated by X. By Proposition 3.5 A is not maximally independent w.r.t H. Further, since $X \models \neg \bigwedge A$, we know that B is not maximally independent w.r.t. H for any $B \supseteq A$, and so in fact A is not independent w.r.t. H. Then there must be some edge E of H such that $E \subseteq A$. By construction of H there must be some $p \in L_n$ and witness set W for $\neg p$ for which $E = W \cup \{p\}$, and so the claim is proved.

$\Leftarrow$ Trivial.

The second part of the corollary is also clear. Indeed, A must contain $W \cup \{p\}$, but if $A \supset W \cup \{p\}$ then there is a proper subset B of A for which $X \models \neg \bigwedge B$, namely $B = W \cup \{p\}$. □

Now we are ready to make the link between normal positive frames and hypergraphs. See Figure 3.2 for an example of how we can use this correspondence to give graphical representations of normal positive frames.

Theorem 3.7 *There is a 1-1 correspondence between normal positive frames and reflexive sets of classification sentences.*

Proof. Denote by f the mapping that takes a reflexive set of classification sentences to its corresponding normal positive frame. By definition there corresponds at least one reflexive set of classification sentences to each normal positive frame and so the mapping f is trivially surjective. To prove injectivity consider two reflexive sets X, Y of classification sentences on L_n, and suppose that they both determine the same normal positive frame T; that is $f(X) = f(Y) = T$. Notice then that, by definition of T, for any sentence $\theta \in SL_n$, $X \models \theta \Leftrightarrow Y \models \theta$. Let

$$X = \left\{ \neg p_i \leftrightarrow \bigvee_{j=1}^{k_i} \bigwedge W_{i,j} \mid 1 \leq i \leq n \right\}$$

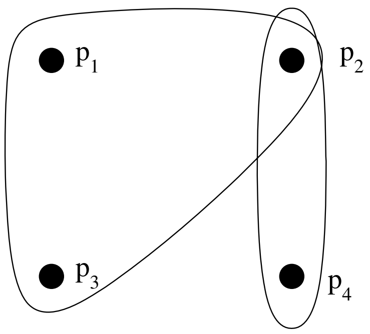
and

$$Y = \left\{ \neg p_i \leftrightarrow \bigvee_{j=1}^{\kappa_i} \bigwedge V_{i,j} \mid 1 \leq i \leq n \right\}.$$

Consider some $W_{i,j}$. By Corollary 3.6 $X \models \neg \bigwedge (W_{i,j} \cup \{p_i\})$, and for any $B \subset W_{i,j} \cup \{p_i\}$, $X \not\models \neg \bigwedge B$.

Of course then these same properties hold for Y. Again by Corollary 3.6 we can see that there is some p_l and witness set $V_{l,m}$ for which $W_{i,j} \cup \{p_i\} = V_{l,m} \cup \{p_l\}$.

Now if $l \neq i$ then $p_i \in V_{l,m}$, and since Y is a reflexive set there must be some witness set $V_{i,j'}$ such that $V_{i,j'} \cup \{p_i\} = V_{l,m} \cup \{p_l\}$. If $l = i$ we can



This hypergraph is equivalent to the normal positive frame defined by

$$\neg p_1 \leftrightarrow (p_2 \wedge p_3)$$
$$\neg p_2 \leftrightarrow (p_1 \wedge p_3) \vee p_4$$
$$\neg p_3 \leftrightarrow (p_1 \wedge p_2)$$
$$\neg p_4 \leftrightarrow p_2$$

Figure 3.1: Correspondence between normal positive frames and hypergraphs

just set $j' = m$.

Hence for every $W_{i,j}$ there is some $V_{i,j'} = W_{i,j}$. A similar argument provides the converse relationship, and so we see that up to a reordering of the witness sets $X = Y$. Hence the mapping f is indeed injective. □

Corollary 3.8 *There is a 1-1 correspondence between normal positive frames and simple hypergraphs of anti-rank 2.*

Proof. Immediate from Theorem 3.4 and Theorem 3.7. □

3.2.1 Positive Frames and Hypergraph Results

In this section we present some results which rely on the correspondence defined by Corollary 3.8 for their proof.

Firstly, we see that contingency is a property of all normal positive frames. In other words, the fairly natural constraints (of symmetry and efficiency of representation — see Definition 3.6) we imposed on positive frames to make them "normal" have contingency as a consequence.

Proposition 3.9 *Normal positive frames are consistent and contingent.*

Proof. Let T be a normal positive frame on L_n and let H be the hypergraph generated by T. Since H is of anti-rank 2 there must exist non-empty independent sets of H, and hence there must exist non-empty maximal independent sets of H. Hence by Proposition 3.5 there is at least one valuation on L_n which makes T true, and so T is consistent.

For contingency, consider $p \in L_n$. Since every edge of H has cardinality at least 2, $\{p\}$ is independent w.r.t. H and so there exists at least one maximal independent set of H which contains p. Again by Proposition 3.5 then there exists at least one valuation v consistent with T for which $v(p) = 1$ and hence $T \not\models \neg p$.

Finally, to prove $T \not\models p$ take any witness set W for $\neg p$. Since W does not contain any edge of H then by Corollary 3.6 then $T \not\models \neg \bigwedge W$. Then since $T \models \bigwedge W \to \neg p$ we have $T \not\models p$, and so T is contingent as required. □

Our final result in this shows that normal positive frames are the weakest contingent positive frames — that is, "normality" is a stronger condition than contingency.

Proposition 3.10 *If T is a contingent positive frame on L_n there is a normal positive frame T^* on L_n s.t. $T \models T^*$.*

Proof. Take a contingent positive frame T on L_n. Then for $1 \leq i \leq n$ there are $k_i \geq 1$ distinct witness sets $\emptyset \neq W_{i,1}, \ldots, W_{i,k_i} \subseteq \backslash\{p_i\}$ s.t.

$$T \models \left\{ \neg p_i \leftrightarrow \bigvee_{j=1}^{k_i} \bigwedge W_{i,j} \;\middle|\; 1 \leq i \leq n \right\}.$$

For each i, j set $D_{i,j} = W_{i,j} \cup \{p_i\}$. Then put

$$H = \left\{ \ D_{i,j} \ \middle| \ 1 \leq i \leq n, \ 1 \leq j \leq k_i \ \right\}$$

and let H^+ be H with all repetitions deleted.

Now remove every $F \in H^+$ for which $\exists \ E \in H^+$ s.t. $E \subsetneq F$ and call the resulting family H^*. That is,

$$H^* = H^+ \ \left\{ \ F \in H^+ \mid \exists E \in H^+, E \subsetneq F \ \right\} .$$

Claim: H^* is a simple hypergraph of anti-rank 2.

Clearly, for all $E \in H^*$, $|E| \geq 2$ since $E = W_{i,j} \cup \{p_i\}$ for some i, j and so H^* is clearly of anti-rank 2. Now, suppose that

$$\bigcup_{E \in H^*} E \neq L_n .$$

Then there is some $p_i \in L_n$ s.t. $p_i \notin \bigcup_{E \in H^*} E$. But

$$\begin{aligned} & \bigcup_{E \in H^+} E = L_n \\ \Rightarrow \quad & \forall \ F \in H^+ \text{ containing } p_i, \ \exists \ E \in H^+ \text{ s.t. } E \subseteq F \setminus \{p_i\} \\ \Rightarrow \quad & \forall \ 1 \leq j \leq k_i, \ \exists \ E \in H^+ \text{ s.t. } E \subsetneq W_{i,j} . \end{aligned}$$

Now $E \in H^+$ gives $T \models \neg \bigwedge E$ and so for for each $1 \leq j \leq k_i$ we have

$$T \models \neg \bigwedge W_{i,j} .$$

But since $T \models \neg p_i \leftrightarrow \bigvee_{j=1}^{k_i} \bigwedge W_{i,j}$ then

$$\begin{aligned} & T \models \neg p_i \leftrightarrow \bigvee_{j=1}^{k_i} \bot \\ \Rightarrow \qquad & T \models p_i \end{aligned}$$

which is a contradiction to the contingency of T. Hence $\bigcup_{E \in H^*} E = L_n$, and so H^* is indeed a hypergraph of anti-rank 2. It is clearly simple by its construction from H^+, and so the claim is proved.

So by Corollary 3.8 there is a normal positive frame T^* corresponding to H^*. Suppose T^* is

$$T^* = \left\{ \neg p_i \leftrightarrow \bigvee_{j=1}^{k'_i} \bigwedge W'_{i,j} \mid 1 \leq i \leq n,\ \emptyset \neq W'_{i,j} \subseteq L_n \setminus \{p_i\} \right\}.$$

It remains to show that $T \models T^*$. First note that since T is contingent it is also consistent. Take some $\alpha \in \mathrm{At}(T)$ and consider each p_i. There are two cases:

1. $\boldsymbol{\alpha \models p_i}$ Suppose there is some $W'_{i,j}$ s.t. $\alpha \models p_i \wedge \bigwedge W'_{i,j}$. Then by construction of T^* there is some $E \in H^*$ s.t. $E = \{p_i\} \cup W'_{i,j}$. Similarly, by construction of H^* there is some $p_{i'} \in E$ and j' s.t. $W_{i',j'} = E \backslash \{p_{i'}\}$. Thus we have

$$\alpha \models p_{i'} \wedge \bigwedge W_{i',j'}$$

which contradicts $\alpha \models \neg p_{i'} \leftrightarrow \bigvee_{j=1}^{k_{i'}} \bigwedge W_{i',j'}$. Hence

$$\alpha \models p_i \to \neg \bigvee_{j=1}^{k'_i} \bigwedge W'_{i,j}\,. \tag{3.1}$$

2. $\boldsymbol{\alpha \models \neg p_i}$

   Since $\alpha \models T$ then there is a witness set $W_{i,j}$ s.t. $\alpha \models \neg p_i \wedge \bigwedge W_{i,j}$. By construction of H^* there is $E \in H^*$ s.t. $E = W_{i,j} \cup \{p_i\}$ and by construction of T^* there is some $1 \leq j' \leq k'_i$ s.t. $W'_{i,j'} = W_{i,j}$. Therefore

$$\begin{aligned} & \alpha \models \neg p_i \wedge \bigwedge W'_{i,j'} \\ \Rightarrow \qquad & \alpha \models \neg p_i \rightarrow \bigvee_{j=1}^{k'_i} \bigwedge W'_{i,j} \end{aligned} \tag{3.2}$$

So for each $1 \leq i \leq n$ and $\alpha \in \text{At}(T)$ we have from (3.1) and (3.2)

$$\alpha \models \neg p_i \leftrightarrow \bigvee_{j=1}^{k'_i} \bigwedge W'_{i,j}$$

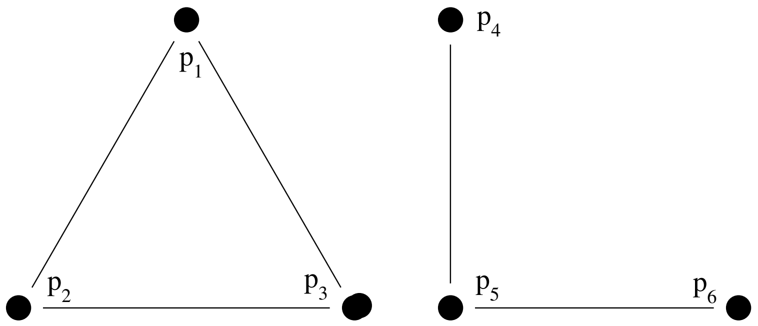
and so $T \models T^*$. □

3.3 1-frames

In Definition 2.1 we allowed unnamed conjunctions of observable properties to be witnesses to the negation of other properties. In this section now deny this possibility, and insist upon the stronger principle that the negation of a property P can only be asserted by the assertion of a witness property N which is incompatible with P. translating this into the terminology of this chapter we get the following definition:

Definition 3.11 (1-frames).
A **1-frame** is a positive frame T for which $|W_{i,j}| = 1$ for every witness set $W_{i,j}$.

Notice that an immediate consequence of Corollary 3.8 is that there is a 1-1 correspondence between normal 1-frames and simple undirected graphs with no isolated vertices. This gives us a very nice visual representation of the "incompatibility structure" of normal 1-frames. Two nodes in the graph corresponding to a normal 1-frame T are connected by an edge if and only iff their respective propositions are mutually incompatible in T. For example,



This graph is equivalent to the normal 1-frame defined by

$$\begin{array}{ll} \neg p_1 \leftrightarrow p_2 \vee p_3 & \neg p_2 \leftrightarrow p_1 \vee p_3 \\ \neg p_3 \leftrightarrow p_1 \vee p_2 & \neg p_4 \leftrightarrow p_5 \\ \neg p_5 \leftrightarrow p_4 \vee p_6 & \neg p_6 \leftrightarrow p_5 \end{array}$$

Figure 3.2: Correspondence between normal 1-frames and simple graphs

3.3.1 Interpretation of 1-frames

There seems to be a sense in which 1-frames are more "natural" than positive frames as a model of the perceptual process described in Section 2.2. We outlined in that section a theory of what it means to directly perceive a property or an attribute. There is an argument to be made then, that unless a particular conjunction of observable properties is itself a property, then it is not an observable property itself — rather it is one which must be inferred. This approach to the process of perception and subsequent inference appears to be a much closer model of the thesis outlined in Chapter 2 in that the perception part of the process is much simpler.

For this reason, we will later examine 1-frames as a model for uncertain reasoning.

3.3.2 Some results from graph theory

In Chapter 6 we will study an inference process which uses normal 1-frames as its starting point. The correspondence between normal 1-frames and simple graphs will become very useful for this purpose, especially the correspondence between the maximal independent sets of a graph and the valuations of the corresponding 1-frame. In particular, the maximum possible number of maximal independent sets is important. This is a problem solved for simple graphs by Moon and Moser[2] in [32], and so give the following definition:

Definition 3.12 (Moon and Moser function).
The maximum number of maximal independent sets of a graph of order n is denoted $m\,(n)$ and is called the **Moon and Moser function**.

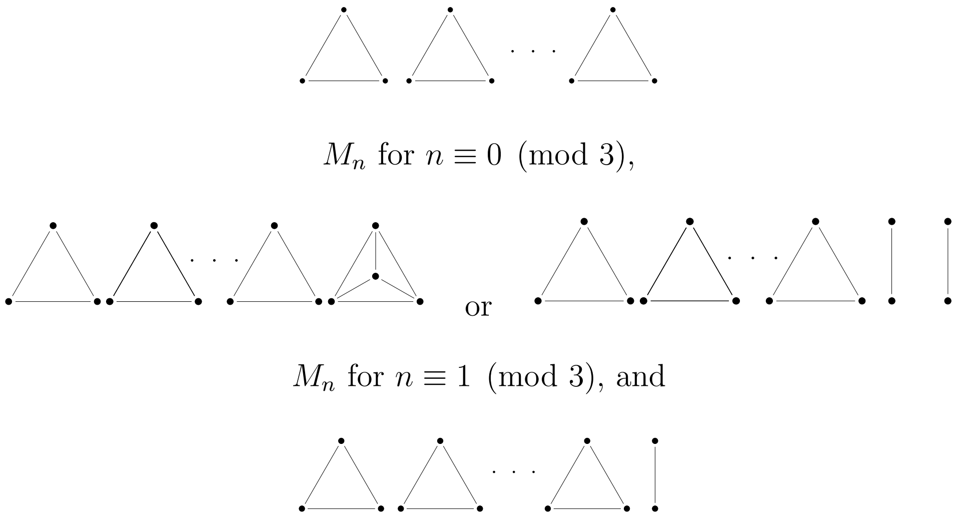
The following theorem from [32] gives the value of $m\,(n)$ and the graphs on which this maximum is attained:

Theorem 3.11 (Moon and Moser [32]) *If $n \geq 2$ then the largest possible number $m\,(n)$ of maximal independent sets for a graph on n vertices is given by*

$$m\,(n) = \begin{cases} 3^{n/3} & \textit{if } n \equiv 0 \pmod 3 \\ 4.3^{[n/3]-1} & \textit{if } n \equiv 1 \pmod 3 \\ 2.3^{[n/3]} & \textit{if } n \equiv 2 \pmod 3 \end{cases}$$

[For completeness and convenience we will also set $m\,(1) = 1$]

[2] We note that the same problem was also solved by Erdös in [11], but the Moon & Moser solution also gives a constructive method for finding the graphs which realise the maximum.



M_n for $n \equiv 0 \pmod{3}$,

M_n for $n \equiv 1 \pmod{3}$, and

M_n for $n \equiv 2 \pmod{3}$.

Figure 3.3: The Moon & Moser Graphs for $n > 2$

Furthermore, for $n \geq 2$ these values of $m(n)$ are attained by the graphs denoted by M_n and shown in Figure 3.11.

We will also find it useful to be able to count the maximum number of maximal independent sets in a *connected* graph. This was calculated by Griggs *et al.* in [17]. We define

Definition 3.13 (Griggs Function).
The maximum number of maximal independent sets of a connected graph of order n is denoted $g(n)$ and is called the **Griggs function**.

Theorem 3.12 (Griggs et al [17]) *The maximum number $g(n)$ of maximal independent sets of a connected graph on n vertices is given by*

$$g(n) = n$$

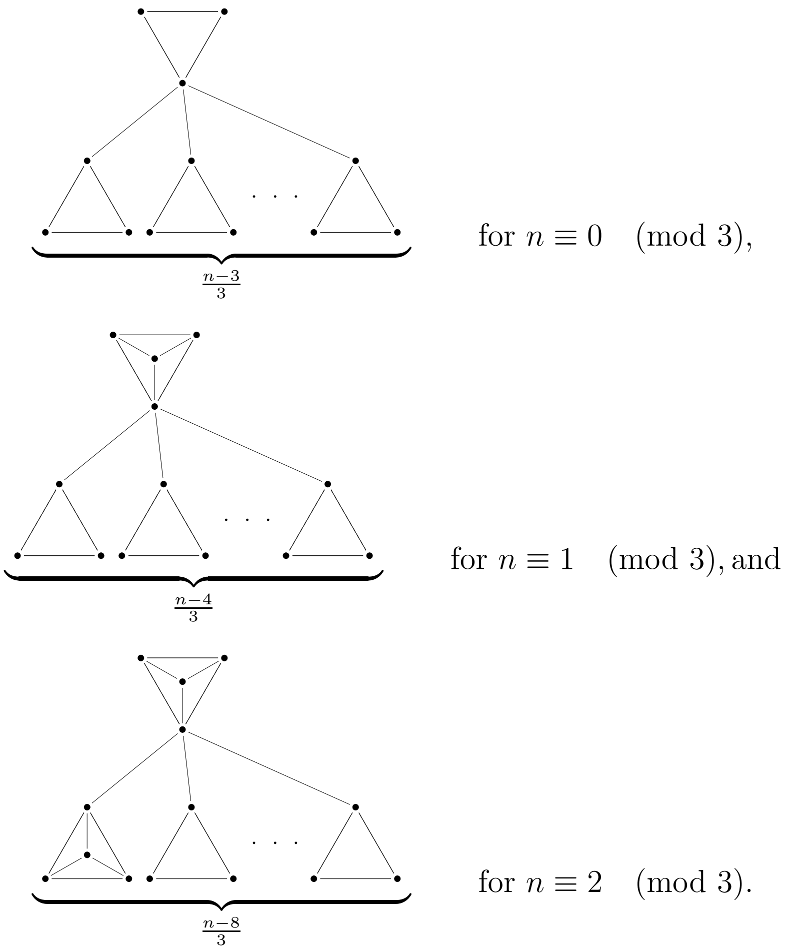

$\frac{n-3}{3}$
for $n \equiv 0 \pmod 3$,
$\frac{n-4}{3}$
for $n \equiv 1 \pmod 3$, and
$\frac{n-8}{3}$
for $n \equiv 2 \pmod 3$.

Figure 3.4: The Griggs graphs for $n \geq 6$

for $n \leq 5$*, and*

$$g(n) = \begin{cases} 2.3^{\frac{n-3}{3}} + 2^{\frac{n-3}{3}} & \textit{if } n \equiv 0 \mod 3 \\ 3^{\frac{n-1}{3}} + 2^{\frac{n-4}{3}} & \textit{if } n \equiv 1 \mod 3 \\ 4.3^{\frac{n-5}{3}} + 3.2^{\frac{n-8}{3}} & \textit{if } n \equiv 2 \mod 3 \end{cases}$$

for $n \geq 6$*. As for* $m(n)$*,* $g(0) = 1$*.*

Furthermore, for $n \geq 6$ *the extremal graphs* E_n *which realise these values of* $g(n)$ *are shown in Figure 3.12.*

For $n \leq 4$ *we have* $E_n = K_n$*, the complete graphs on* n *vertices. For* $n = 5$

however, there are four extremal graphs E_5. One of these is C_5, the circuit on 5 *vertices, and the other three are shown in Figure 3.5.*

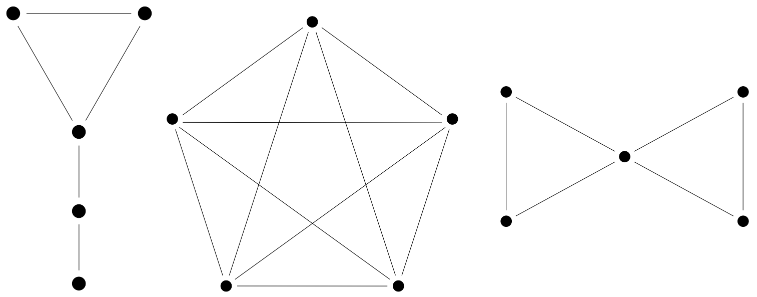


Figure 3.5: The Griggs graphs E_5 for $n = 5$

3.4 Conjunctively Closed Frames

we turn now to the second principle proposed in Section 2.2.4. Definition 2.2 gives the Principle of Conjunctive Closure as the prescription that any consistent conjunction of observable properties must be an observable property itself. In other words, that conjunction should have a name in our logical structure. In our propositional logic framework this can be expressed by the following definition:

Definition 3.14 (Conjunctively Closed Frames).
A positive frame T on a propositional language L is called **conjunctively closed** if for all $X \subseteq L$ s.t. $T \not\models \neg \bigwedge X$ there is some $p \in L$ s.t.

$$T \models p \leftrightarrow \bigwedge X$$

In other words, for every set X of L consistent with T there is some p in L which is T-equivalent to the conjunction of X. For brevity's sake we will write "c-frame" as an abbreviation for Conjunctively Closed Frame throughout the remainder of this thesis, and we will use the notation $\overline{T}$ to indicate that a positive frame T is a c-frame.

Notice that for any contingent c-frame T there is a conjunctively closed 1-frame equivalent to T formed by the removing all witness sets W which are inconsistent with T and then replacing all remaining witness sets W with their corresponding names. Hence the remarks on 1-frames in Section 3.3.1 also apply.

C-frames therefore appear to be the “best” model of the perceptual process that we have, in the sense that they most closely capture the notions involved in our discussions of perception in Chapter 2. In the next chapter we will use them to define an inference process.

Chapter 4

Inference Processes & Positive Frames

This chapter begins with a brief discussion of probability functions and inference processes together with some of the mathematical definitions used later in the chapter. We then define an inference process combining arguments from maximum entropy with the c-frames developed in the previous chapter. This is followed by two justifications for the use of this inference process, which present characterisations of it as a model of expert reasoning based upon a large experience base.

4.1 Probabilistic Inference

This section describes some of the basic definitions required to develop a probabilistic approach to uncertain reasoning with positive frames. We describe the notation we use to work with probability functions, and what it means for a probability function to be consistent with a positive frame. We then give a brief discussion of what is meant by "probabilistic inference" before setting out some technical details necessary to define and analyse the inference process set out in Section 4.2.

Definition 4.1 (Probability functions).

We follow [34] (pages 10–14) in defining probability functions on propositional languages. That is, given a finite propositional language L, $w : SL \to [0,1]$ is a **probability function** on L if, for all $\theta, \phi \in SL_n$ w satisfies the two axioms

(P1) If $\models \theta$ then $w(\theta) = 1$

(P2) If $\models \neg(\theta \wedge \phi)$ then $w(\theta \vee \phi) = w(\theta) + w(\phi)$

An important representation theorem is also given in [34] which defines a correspondence between a probability function on L and the value it gives to the atoms of L. As such it is possible to identify a probability function w with the vector

$$\langle w(\alpha_1), w(\alpha_2), \ldots, w(\alpha_J)\rangle \in \mathbb{D}^L = \left\{ \vec{x} \in \mathbb{R}^J \;\middle|\; \vec{x} \geq 0;\ \sum_{i=1}^{J} x_i = 1 \right\}$$

where $\alpha_1, \ldots, \alpha_J$ run through all atoms of L. $\mathbb{D}^L$ here denotes the set of all vectors corresponding to probability functions on L. If we have $L = L_n$ rather than $\mathbb{D}^{L_n}$, we will often write $\mathbb{D}^n$.

It will be necessary for us to consider probability functions over different size languages which are in some sense 'the same.' To this end for a probability function w on L_k, for any $n \leq k$ we denote the **restriction** of w to L_n as $w \restriction L_n$, and define it on the atoms of L_n by

$$w \restriction L_n(\alpha) = \sum_{\beta \in \mathrm{At}(L_k);\ \beta \models \alpha} w(\beta)$$

for all $\alpha \in \mathrm{At}(L_n)$.

We will be interested in studying probability functions which correspond to positive frames. The natural way in which a probability function might be said to 'correspond to' a logical theory is if it gives probability 1 to any sentence which is true relative to that theory. That is, for a positive frame T on L, w is a probability function for T iff $w(\theta) = 1$ for each $\theta \in SL$ s.t. $T \models \theta$. Notice then that no probability function can correspond to an inconsistent theory, since that would give $w(\theta) = w(\neg\theta) = 1$, which contradicts (P1) and (P2) above.

Note that it is sufficient to check that the above condition holds for any set $X \subseteq SL$ s.t. $X \equiv T$, rather than checking for every sentence of SL. So for example, if T is a normal positive frame, w is a probability function for T iff $w(\theta) = 1$ for all θ in the RCS corresponding to T (Recall that there is exactly one such RCS by Theorem 3.7).

An immediate consequence of this definition is that a probability function w corresponds to a general positive frame T iff we have $\sum w(\alpha) = 1$ where the sum is over all α in $\mathrm{At}(T)$.

Equivalently[1], a probability function w can be said to correspond to a positive frame T on L if it only gives non-zero probability to atoms of L which are in $\mathrm{At}(T)$. That is

$$\alpha \in \mathrm{At}(L) \setminus \mathrm{At}(T) \Rightarrow w(\alpha) = 0$$

4.1.1 Inference Processes, Constraint Sets and Solution Sets

Our aim in this chapter is to present an inference process for finite propositional languages. An inference process is, generally speaking, some rule which, given

[1] It may seem a little excessive to give so many equivalent conditions for an essentially simple concept, but all of them will have occasion to be useful in the course of this thesis so they are given together here for the sake of clarity

some information or data or knowledge about allowable or possible probability functions, selects a subset of those probability functions. Phrased in the terminology of uncertain reasoning and belief functions, inference processes are often portrayed as being a way of picking the "best" belief function available under a certain set of constraints (see for example [7], [8], [19], [30],[34], [35],[36],[37], [38]). What constitutes "best" is a question of some debate. Indeed, in general, for a given set of knowledge K there will be many probability functions consistent with K — often uncountably many[2].

An established approach within uncertain reasoning is to impose certain principles or conditions on our inference process that we think are reasonable or desirable, and to analyse the resulting inference processes. One such principle is the Equivalence Principle (see [34], pp. 82–87), a principle which dates right back to Laplace's work on the founding of modern probabilistic reasoning, yet which was not made explicitly clear until Jeffreys stated it in the 1930's (see [23], p.7). In essence, this principle states that in circumstances where we have the same knowledge, we should assign the same probability function. We will see that the inference processes described here satisfy this principle trivially, and shall study their satisfaction or otherwise of a range of other principles.

Turning to the question of what we intend to denote by information or data or knowledge, we comment that it is a common step in probabilistic uncertain reasoning to formalise our knowledge as a constraint set: that is, some constraints on the possible or allowable probability functions. We would like to think that a constraint set arising from given knowledge (or information or data or beliefs) K determines a set of probability functions which are "consistent" with the information in K.

[2] Assuming of course that K is consistent, in the sense that there is a probability function which agrees with K. This is by no means a trivial assumption, as discussed in [34], [50] and [54]

In general, we would like the set of probability functions determined by our constraint sets to be both closed and convex for reasons of mathematical convenience. Obviously in practise it is not always convenient or even possible to establish such a constraint set[3]; however an analysis of the more mathematically convenient case is one way in which we can advance our understanding of the more complicated situation. In practise we will impose a slightly more rigid restriction on our constraint sets:

Definition 4.2 (Constraint Sets and Solution Sets).

Given a finite propositional language L a constraint set Σ is a finite set of linear constraints on the values assigned to the atoms by a probability function w on L. That is, for $L = L_n$ we have m constraints of the form

$$\sum_{j=1}^{2^n} a_{i,j} w(\alpha_j) \leq b_i$$

where $i = 1, \ldots, m$ and $a_{i,j}, b_i \in \mathbb{R}$. Denote the set of all such constraint sets on L by CL. Notice that these constraints can also be given in matrix form as

$$\mathbf{A}.\vec{w} = \vec{b}$$

where $\mathbf{A}$ is an $m \times 2^n$ matrix given by $\mathbf{A} = (a_{i,j})$ while $\vec{w}$ and $\vec{b}$ are vectors of size 2^n given by $\vec{w} = \langle w(\alpha_1), w(\alpha_2), \ldots, w(\alpha_{2^n}) \rangle$ and $\vec{b} = \langle b_1, b_2, \ldots, b_{2^n} \rangle$.

For a constraint set Σ on L, we denote the set of probability functions on L which satisfy the constraints given in Σ by $V(\Sigma)$. If Σ is given on L_n then for all $k \geq n$ denote by $V_k(\Sigma)$ the set of all probability functions on L_k whose restriction to L_n satisfies the constraints in Σ. Denote by $\mathbb{D}^L(\Sigma)$ and $\mathbb{D}^k(\Sigma)$ the subsets of

[3] See [54] for a discussion of the problems involved and a suggestion of how to establish such sets

$\mathbb{D}^L$ and $\mathbb{D}^k$ corresponding to $V(\Sigma)$ and $V_k(\Sigma)$.

Note that both $V(\Sigma)$ and $\mathbb{D}^L(\Sigma)$ are closed and convex. We turn now to some special types of constraint set that we will be particularly interested in.

A **proper constraint set** $\Sigma \in CL_n$ is one which does not forces the elements of L_n to be contingent — i.e., it does not permit probability functions to insist that any of the elements of L_n are either necessarily true or false. We formalise this as the condition that each $w \in V_n(\Sigma)$ has the property that for all $p_i \in L_n$, $w(p_i) \in (0, 1)$ — that is, $w(p_i) \neq 0$ or 1.

We say that a constraint set $\Sigma \in CL_n$ is called **adamant** (on L_n) if there is some subset $\{p_{i_1}, p_{i_2}, \ldots, p_{i_r}\}$of L_n such that Σ forces the disjunction of this subset to be always true. Formally, for all $w \in V_n(\Sigma)$ we have

$$w(p_{i_1} \vee p_{i_2} \vee \ldots \vee p_{i_n}) = 1$$

Constraint sets which are not adamant are called **ethereal**.

Now that we have made the notion of a probability function being consistent with a constraint set clear, we consider what it means for a (general) positive frame to be consistent with a constraint set. We think of the constraint set Σ embodying our knowledge about a certain language L, say, whilst a positive frame imposes a certain logical structure on that language. Therefore, for a GPF T to be consistent with Σ, the logical structure of T must be realisable under the constraints of Σ. Recalling our remarks on Definition 4.1 then, there must exist some probability function $w \in V(\Sigma)$ s.t. $w(\alpha) = 1$ for all $\alpha \in \mathrm{At}(T)$. In the case that T is on L_n and Σ is defined on L_k, where $k \leq n$, then we say T is consistent with Σ if there is some $w \in V_n(\Sigma)$ such that $w(\alpha) = 1$ for all atoms of T.

For yet further notational convenience we denote by $V(\Sigma, T)$ the set of probability functions consistent with both T and Σ, and let $\mathbb{D}(\Sigma, T)$ denote the corresponding subset of $\mathbb{D}^L$. Note that here we do not need to index these sets with the language size since it must be the same as the size of T.

Finally, now that we are clear on probability functions and constraint sets we can define what we take to be an inference process:

Definition 4.3 (Inference Processes).
An inference process on a propositional language L_n is a function $N : CL_n \to \mathscr{P}_0(\mathbb{D}^n)$ which assigns to every constraint set on L_n a finite set of probability functions which are each consistent with that constraint set.

That is, for each constraint set $\Sigma \in CL_n$, $N(\Sigma)$ is a finite set

$$\{w_1, w_2, \dots, w_r\} \subseteq V_n(\Sigma)$$

4.2 The CFE$_k$ Inference Process

4.2.1 The Concept

We present here an inference process (in the sense defined in the last section) which combines the principle of maximum entropy with the Classification Principle and the Principle of Conjunctive Closure, as discussed in Chapter 2 and given mathematical formalisation in Chapter 3. The concept is to restrict the maximum entropy inference process to c-frames.

As discussed in Chapter 3, I claim that as propositional structures, c-frames capture the principles outlined in Section 2.2.4. I also claim that the principle of maximum entropy should be applied to uncertain reasoning with these structures, for the reasons outlined in Section 2.1.1. The CFE inference process defined in the next section is an application of the principle of maximum entropy with the additional conditions that the Classification Principle and the Principle of Conjunctive Closure also apply.

4.2.2 Formalisation

Definition 4.4 (The Entropy Function).

For a probability function w on L_n let the **indexed entropy** of w be

$$\mathrm{H}_n(w) = \prod_{\alpha \in L_n} w(\alpha)^{-w(\alpha)}$$

Note that this is the exponent of the entropy function as it is usually used and derived elsewhere (for example in [24], Chapter 11) — we use this form of the function purely for convenience with certain of the calculations performed in this thesis. For the sake of conformity, we also define

$$\mathrm{h}_n(w) = -\sum_{\alpha \in L_n} w(\alpha) \ln w(\alpha)$$

which we will also have some use for. Note that $\mathrm{h}_n(w) = \ln \mathrm{H}_n(w)$, and since most of our entropy calculations will be concerned with maximising entropy, and specifically on which probability functions it is maximised, it will usually be irrelevant which form of the function we use in a given case.

Definition 4.5 (Molecular Weight and Σ-minimality).

Suppose that Σ is a constraint set on L_k. The minimum n for which there is a c-frame T on L_n which is consistent with Σ is called the **molecular weight** of Σ and is denoted $\xi(\Sigma)$.

Any c-frame $\overline{T}$ on a language L_n where $n = \xi(\Sigma)$ is called **Σ-minimal**.

Definition 4.6 (The CFE$_k$ Inference Process).

We give here a definition of an indexed inference process CFE$_k$ from the set of adamant constraint sets on L_k to the set $\mathbb{D}^k$ of probability functions on L_k.

Suppose Σ is an adamant constraint set on L_k. Let the Σ-minimal c-frames which are consistent with Σ be denoted $T_1, T_2, \ldots, T_K$.

For each T_i let w_i be that $w \in V_{\xi(\Sigma)}(\Sigma, T_i)$ which maximises the entropy. That is, set

$$w_i = \underset{w \in V_{\xi(\Sigma)}(\Sigma, T_i)}{\operatorname{argmax}} \mathrm{H}_{\xi(\Sigma)}(w)$$

Let $\mathrm{H}_{MAX} = \max\left\{\mathrm{H}_{\xi(\Sigma)}(w_1), \mathrm{H}_{\xi(\Sigma)}(w_2), \ldots, \mathrm{H}_{\xi(\Sigma)}(w_k)\right\}$ and then, finally we can set

$$\mathrm{CFE}_k(\Sigma) = \left\{ w \upharpoonright L_k \;\middle|\; w = w_i \text{ for some } i = 1, \ldots, K \text{ and } \mathrm{H}_{\xi(\Sigma)}(w_i) = \mathrm{H}_{MAX} \right\}$$

4.3 A Simple Characterisation of CFE$_k$

We now move on to providing a characterisation of the CFE$_k$ process in an attempt to justify its use. The general concept is similar to that of Jaynes' justification of Maximum Entropy given in Chapter 11 of [24] and, in a slightly different form, explained in detail by Paris and Vencovská in [35]. Indeed, the justification given here follows that given in [35] quite closely, and uses or adapts many of the results given there. The argument is essentially a version of Boltzmann's derivation of the Maximum Entropy method as discussed in [22].

The idea is simple: we consider 'examples' distributed at random, which in some sense satisfy a set of constraints. In [35] it is shown that when these examples are distributed according to the constraints, then as their number grows without limit the vast majority of them will be near the classical maximum entropy solution. In our case we also impose the restriction that each example corresponds to some c-frame, in a sense which we will make clear later. It is the aim of this section to show that distributing the examples in this way leads to the examples clustering around the (finite number of) CFE$_k$ solutions.

4.3.1 The Model

The concept we use to justify the CFE$_k$ inference process here is similar to that used in [35] — in fact it is the same model, but with an extra condition imposed. Essentially, the idea is that we consider an expert who has a great deal of experience of a certain field. This experience is modelled as a large set of examples of the expert's field, and various properties are identified with subsets of this set. The expert's knowledge, drawn from this experience, is modelled as a set of approximate constraints on the relations between the properties. A typical such approximation may be "About half the examples have property P." The next step in the process is to consider the many ways of assigning the subsets of the example set to the properties in such a way as to conform with the constraints. This is where we differ slightly from Paris and Vencovská in [35]. Whereas there is no restriction on the way in which these subsets are assigned in their paper, here we impose the restriction that each assignment must correspond to a c-frame. There is a natural way to define a probability function from each assignment, and it turns out that as the number of examples grows without bound then so the vast majority of probability functions so defined become arbitrarily close to the solutions of CFE$_k$. The remainder of this section gives a more detailed explanation of this process. It is in the nature of such abstract ideas that they can seem quite vague at times, and so a detailed example is given to illustrate the method described.

Suppose we have an expert in some field whose knowledge we wish to model mathematically — this may be for the purpose of making predictions, or perhaps we simply wish to understand better how they themselves reason from their experience to conclusions. Our hypothetical agent could be expert in locating valuable oil fields for example, a gambler betting on horse racing, or a pathologist studying tumour slides. This last example is indeed the motivation of Paris and Vencovská in [35]. Such an agent is a common starting point for the study and application of intelligent reasoning and "artificial intelligence", and predictive systems arising from their consideration are naturally known as *expert systems*. However, we should note that the nature of the expert, whether an oil prospector, a stockbroker or a pathologist, is not important to our argument here; what is important is the way in which we model their knowledge and experience.

We make a connection between "experience" and "knowledge" here. The field of epistemology is a fascinating area of philosophy[4], but we shall not be overly concerned with the status of our expert's "knowledge" here. Instead,we simply remark that we consider the expert to have a certain amount of information on their field which has been acquired in a number of ways — they will have read a certain amount of literature on their field of expertise; they may have discussed certain phenomena with other experts; they will (ideally) have seen a large number of examples of their field — from which they have formed their opinions and beliefs. Hence *experience* seems a suitable term to describe the many ways in which our expert has acquired his/her information, while we use the term *knowledge* to describe the more general beliefs formed and distilled from this experience. Note that this is not an attempt at defining what "knowledge"

[4] For a stimulating introduction to this field, [49] contains a selection of articles covering many different aspects of the theory of knowledge.

is in the strict epistemological sense, it is simply a convenient terminology for the purposes of this thesis.

As remarked above, we will be following quite closely the model of experience developed in [35]. Paris and Vencovská there illustrate the modelling of what they term *old knowledge* by developing the example of a pathologist visually examining lymphoma/epithelial tumour slides — it is worth noting that this arose from a real attempt to create an expert system to study such slides. A different example is developed in this thesis for two reasons: the first is simply to give the reader a fresh example to consider alongside that set out in [35]. Secondly, I wish to illustrate that this method of modelling experience is applicable to a wide range of expert knowledge — there is nothing special about tumour slides *per se* in this context[5].

For these reasons, the example we give will be one of a gambler whose specific field of expertise is horse racing. Far from being a trivial or frivolous example of uncertain reasoning, this is in fact an ideal field for analysis of how experts come to decisions based on their experience. One of the primary reasons for this is that serious gamblers will have a huge body of experience. A professional gambler will almost certainly have seen (and bet on) thousands, if not tens of thousands, of races. That our expert has experienced a large number of examples of their field is essential to the argument presented here.

Further, the field lends itself well to statistical analysis. There are a multitude of daily, weekly and annual statistics published on the subject of horse racing, ranging from the minimal daily form guides and results services published in tabloid newspapers, through the more detailed analysis in specialist newspapers such as *The Racing Post* to in-depth statistical analysis of results provided by weightier tomes such as *Timeform*.

[5] Other authors have also modelled experience in a similar way; see for example [39] and [4]

Finally, the whole point of gambling is that it is a question of *forecasting* a result. As discussed in various places earlier, one of the major problems in developing real-life expert systems is that of eliciting consistent information from the experts that accurately reflects their state of knowledge and experience. This problem is likely to be reduced in the case of a professional gambler — they are likely to have a fairly shrewd idea of how to represent their knowledge in a useful format. Indeed, the vast number of books touting "racing systems" available in any moderately sized bookshop is testament to this fact. A superb and practical discussion of "The Horse Racing Problem" is given in [31], especially the first and second chapters. Many professional gamblers do indeed use such systems.

Of course, it may be argued that the fact that our expert is likely to have given such thought as to how to represent their knowledge is likely to distort such a representation. For example, an expert who has previously worked with a particular system, such as one of the rule-based systems presented in [31], may have forced the representation of his/her experience into a format consistent with the rules of the system — and as such does not accurately reflect their experience. This may introduce a distorting factor into the representation of their experience that we can elicit from them, if they are "set in their ways" of representing knowledge about the field of horse-racing. The fact that many professional gamblers do indeed use such systems, and are likely to have considered or used a large number of them over time lends weight to this criticism.

The counter-argument to this point is that the effect of using such a system, presumably with a conscious or sub-conscious evaluation of the success of using the system, will be present in the statements that our expert gives us when we are forming our representation of their experience. That is, *all* of our expert's experience of their field, including such indirect experience as using "racing systems," is contained with the statements that we elicit from them. Indeed, it can be argued that the fact that use of a previous system may have distorted the statements given to us by our expert is not in fact a distortion, but is essential to these statements containing all of the expert's experience and knowledge. In other words, this is the *Watts Assumption* that our knowledge base contains *all* of our expert's experience, not just some of it (See [34]).

Therefore we can be fairly certain that our expert has access to a great deal of experience, both direct and indirect, about their chosen field, and is reasonably likely to be able to present us with a useful representation of their knowledge.

There is one final criticism to be levelled at the choice of a horse racing expert as a useful example of an "expert", and it is an important one. That is simply that most gamblers are not very good! Indeed, the very existence of bookmakers, together with their substantial profits, illustrates this fact. What sense can be found in an analysis of such inadequate reasoning? The answer is twofold: firstly, a practical rebuttal is possible. While it is true that the vast majority of gamblers do lose in the long run, this is partly as a result of the lop-sided books that bookmakers keep. The average bookie keeps his book "over-round" to the tune of 20–30% — that is, if horses were picked at random a loss of 30p could be expected for every £1 staked. Compare this to roulette which operates at 3% over-round [31]. This is an important point — bookmakers *must* increase their margins to turn a profit as the average punter does better than simply picking horses at random, so the gambler as expert is not such an inadequate choice as it initially appears. Further, there are of course professional gamblers, who make a living out of "beating the bookies," turning that 30% loss into a profit. These are the people we should consider an expert in this field.

The second answer to the 'problem' of gambler-as-expert is that the success or otherwise of our expert is, to a certain extent, irrelevant. We are simply trying to model how the expert might evaluate certain probabilities based on their experience. Whether the probabilities so evaluated are profitable to our gambler is not the issue here — we are concerned with the effectiveness or otherwise of our *modelling of the expert's reasoning*, not the success of that reasoning.

Now, there are many factors involved in choosing a horse to bet on. Racing systems vary in their complexity from those that rely on only a few factors such as the *Topspeed* or *Postmark* speed ratings of a horse to those that take in a vast array of factors such as the horse's form, the jockey, the trainer, the weather, the number of days since the horse ran, and so on and on. In this example we shall consider only a few factors — we are not attempting to present a working forecasting system, simply to illustrate how our expert's knowledge can indeed correspond to the mathematical analysis we will give in the next section. Of course, being a gambler our expert is interested in which horse should be expected to win a race under consideration. This is however, far too complicated a problem to serve as a sufficiently illuminating example here. Therefore we will present here an example of how a gambler might judge two important factors which will influence their decision on whether to consider a particular horse as a valid betting proposition, with the information gleaned from such a judgement used in a further decision on whether to bet on a horse or not. These factors will be the horse's standard of fitness and whether it can be judged to be 'trying.' The first factor is a fairly obvious choice — an unfit horse is unlikely to win a race. The second factor however is slightly different. While it would seem obvious that every horse will try to win every race, in reality this is not always the case. Trainers may enter horses into races without the intention of winning for a variety of reasons: to increase its fitness levels, to improve a young horse's experience of racing, or possibly even to lower the handicap for a later race which the trainer has set his/her sights on winning.

Unlike many of the statistical data available to race followers such as form, handicap, distance, etc., these are two factors which are relatively difficult to judge before a race has been run, but will usually be apparent to the experienced gambler after the race. This makes these factors ideal for the purposes of this illustration: the experience the expert has of past races will include which horses were fit and/or trying, but this data is not apparent for the next race to be run. They are however factors which our expert may well need to consider, and a certain number of other factors, of which our expert can have knowledge, may influence the expert's decision on whether the horse is fit and/or trying. For example, it seems likely that our expert may use a simple rule-based system to rule out betting on horses which are either not fit or not trying. Unfortunately, unlike the horse's form, the information on these factors is not available in the newspapers and so our expert must make a judgement based on what data is available.

Let us assume that our expert has given us the following information:

$$\text{Most horses which have run recently are fit} \tag{4.1a}$$

$$\text{Top class trainers usually produce fit horses} \tag{4.1b}$$

where the expert has given us some indication of how to judge who counts as a top class trainer in their opinion, perhaps a list of the 'elite.' The next statement might be about the fact that 'not trying' does occur, but comparatively rarely:

$$\text{The majority of horses try to win} \tag{4.1c}$$

This is obviously not a huge amount of information to go on for our putative system — most obviously the question of a whether or not a horse is trying is still a bit vague. We have information about the circumstances which affect whether a horse is trying or not. Suppose we ask our expert to clarify and receive the following statement

For races classed as Group 1, the horse is almost certainly trying to win (4.1d)

We now have a (small) set of statements from our expert about horse-racing. We would like to rephrase these in a useful fashion. As discussed earlier, we are taking the approach of considering a large set of examples, in this case the runners that the gambler has previously seen or has knowledge of. Let this set be H. We would like to attach some mathematical meaning to phrases like "most... are ...", "usually" and "almost certainly." There are numerous ways of doing this, for example the verbal-numerical scale proposed in [54]. However, for the sake of simplicity let us ask our expert to simply give approximate frequencies to these statements, based on the experience contained in H. We would be led to a set of statements such as

Approximately 60% of horses which have run recently are fit (4.2a)

About 70% of horses produced by top trainers are fit (4.2b)

Around 75% of horses try to win (4.2c)

About 95% of horses try to win Group 1 races (4.2d)

Now we identify (unspecified) subsets of H with the different properties of runners described in the statements as follows. Let

The set of runners which were fit be F

The set of runners which ran recently be R

The set of runners which are trying be T

The set of runners with a top class trainer be C

The set of runners which ran in Group 1 races be G

Now, if for the time being we express "about" and "approximately" as $\approx$, leaving the exact definition until later, we can express (4.2a)–(4.2d) as

$$|R \cap F| \approx \ .6|R| \tag{4.3a}$$

$$|C \cap F| \approx \ .7|C| \tag{4.3b}$$

$$|T| \approx \ .75|H| \tag{4.3c}$$

$$|G \cap T| \approx \ .95|G| \tag{4.3d}$$

We now have a set of approximate constraints on the subsets of H which correspond to our named properties. That is, however we assign the elements of H to the various subsets F, R, T, C and G, this assignment must conform to these constraints — with the additional constraint that the assignment must also correspond to some c-frame, where the precise meaning of such a correspondence is defined in the next section.

Notice that each assignment naturally defines a probability function on the propositional language $L = \{F, R, T, C, G\}$. Suppose we have an assignment $A : L \to \mathscr{P}(H)$ which 'satisfies' our approximate constraints and also corresponds to a c-frame. Each $\alpha \in \mathrm{At}(L)$ will be of the form $F^{\epsilon_1} \wedge R^{\epsilon_2} \wedge T^{\epsilon_3} \wedge C^{\epsilon_4} \wedge G^{\epsilon_5}$ where $\epsilon_i \in \{0, 1\}$. Define a probability function a from A by

$$a(\alpha) = \frac{|A(F)^{\epsilon_1} \cap A(R)^{\epsilon_2} \cap A(T)^{\epsilon_3} \cap A(C)^{\epsilon_4} \cap A(G)^{\epsilon_5}|}{|H|}$$

where for any $X \subseteq H$ we set $X^0 = H \setminus X$ and $X^1 = X$.

The next step is to consider all possible assignments of H. In general, for large H there will be a vast number of such assignments. As we have no more information about the assignments than that contained in the constraints, together with the additional condition that they should conform to c-frames, it seems reasonable to insist that we should take the most ‘common’ assignments as being representative of the expert’s knowledge. In other words, while we do not know exactly which of the many possible assignments of H corresponds to our expert’s actual experience, if there is a class of assignments which constitutes the majority of such assignments, it seems reasonable to assume that our expert’s experience is one of this class. We will show in the next section that majority of the probability functions arising from these assignments will tend to cluster around the CFE solutions to the constraints when H is sufficiently large.

4.3.2 A Mathematical Description

We first give a general outline of the mathematical argument presented in this section. We begin with some preliminary technical definitions such as what we take to be a constraint set, and the precise meaning that we give to $\approx$. We then state our characterisation theorem, and follow it with a proof. To give a brief description of the proof: first we restrict our attention to a single c-frame and apply Theorems[6] 1 and 2 of [35] to show that assignments of examples which

[6] These theorems are stated in Appendix A as Theorem A.2 and Theorem A.3

correspond to a particular c-frame cluster around the maximum entropy solution on that c-frame. This is a fairly standard argument in maximum entropy but will require some careful analysis of technical details to do with how we impose the condition of "corresponding with a c-frame."

We then present a number of Lemmas to illustrate that as the number of examples grows without bound then the assignments of the examples will tend to cluster around a finite set of c-frames. These will be shown to be those c-frames whose probability functions (as selected in the first step) have greatest entropy — in other words the CFE$_k$ solutions to the constraint set. Between them these Lemmas essentially amount to a reworking of the proof of Theorem 1 of [35] to take into account the fact that we are dealing with both the restriction of the argument given there to c-frames and the fact that in general we will have multiple solutions in the case of CFE$_k$, unlike the unique solution given by classical maximum entropy.

Definition 4.7 (Examples and Properties).

We define an **example set** to be simply a non-empty set E of cardinality N. Properties $P_1, \ldots, P_k$ $(k > 0)$ defined on E are subsets of E. We identify

$$\begin{aligned} \neg P_i &\quad \text{with} \quad E \setminus P_i \\ P_i \wedge P_j &\quad \text{with} \quad P_i \cap P_j \end{aligned}$$

Let the set of example sets of size N with k properties be denoted $EX(N,k)$.

From such an example set E we name the **atomic properties** of E as $A_1, A_2, \ldots, A_{2^k}$ and are defined to be the sets of the form

$$P_1^{\epsilon_1} \wedge P_2^{\epsilon_2} \wedge \cdots \wedge P_k^{\epsilon_k}$$

where $\epsilon_i \in \{0,1\}$, $P_i^0 = \neg P_i$ and $P_i^1 = P_i$. Let the set of atomic properties of an example set E be denoted $\mathrm{At}(E)$.

From this definition it is easy to take the constraints given to us by our expert, as in the derivation in the previous section, and present them in a cardinal form:

Definition 4.8 (Expert Constraint Sets).

Let E be an example set of size N with properties $P_1, \ldots, P_k$ defining atomic properties $A_1, \ldots, A_{2^k}$. An **expert constraint set** on E is a set of m constraints, each of which is either of the form

$$\sum_{i=1}^{2^k} x_{i,j}|A_i| \approx c_j N \tag{4.4}$$

$$\text{or} \quad \sum_{i=1}^{2^k} y_{i,j}|A_i| \approx 0 \tag{4.5}$$

where $x_{i,j} \in \{0,1\}$, $y_{i,j} \in \mathbb{R}$ and $c_j \in (0,1)$.

Define the set of expert constraint sets on example sets of size N with k properties $P_1, \ldots, P_k$ to be $EL(N,k)$.

Note that since the $A_1, \ldots, A_{2^k}$ form a partition of E we also have the (mandatory) constraint that

$$\sum_{i=1}^{2^k} |A_i| = N \tag{4.6}$$

and of course for each i we also have

$$|A_i| \geq 0 \tag{4.7}$$

Definition 4.9 (Approximation Relation).

Suppose For N and m as given we now define the relationship $\approx$ on integers

$X, Y \leq N$ as

$$|X - Y| \leq 2^k\sqrt{m} \quad \Rightarrow X \approx Y \tag{4.8}$$

and

$$X \approx Y \quad \Rightarrow \quad |X - Y| \leq \epsilon N \tag{4.9}$$

for some $\epsilon > 0$. Obviously these relationships require $2^k\sqrt{m} \leq \epsilon N$.

A brief word on the $\approx$ relationship: the two relationships which specify $\approx$ are indeed reasonable ones when N is large. (4.8) states that if X and Y differ by less than a specified and constant absolute amount (i.e., $2^k\sqrt{m}$), then they should be considered approximately equal — this makes sense if N is large compared to this value. (4.9) states that if X and Y are approximately equal then they differ by less than a specified proportion of N. Again this seems reasonable. Strictly speaking $\approx$ should be indexed by ϵ, N, k and m, but since this would make our notation cumbersome and almost certainly illegible we shall restrict ourselves to using $\approx$ and commenting on its dependence on these parameters where appropriate.

Set $z_i = \frac{|A_i|}{N}$ for each $i = 1, 2, \ldots, 2^k$. Then (4.4)-(4.7) become

$$\sum_{i=1}^{2^k} x_{i,j} z_i N \approx c_j N \tag{4.10}$$

$$\sum_{i=1}^{2^k} y_{i,j} z_i N \approx 0 \tag{4.11}$$

$$\sum_{i=1}^{2^k} z_i N = N \tag{4.12}$$

$$z_i N \geq 0 \tag{4.13}$$

Notice here that each vector $\mathbf{Z} = \langle z_1, z_2, \ldots, z_{2^k} \rangle$ where $\sum_{i=1}^{2^k} z_i = N$ corresponds to

$$\frac{N!}{\prod_{i=1}^{2^k} z_i!}$$

assignments of the elements of E to the partition formed by $A_1, A_2, \ldots, A_{2^k}$. The following notation will aid in the clarifying the meaning of this thesis:

Notation: If $\mathbf{X}$ is a vector of size 2^n whose entries are all non-negative integers which sum to N, we define

$$\begin{bmatrix} N \\ \mathbf{X} \end{bmatrix} = \frac{N!}{\prod_{i=1}^{2^n} X_i!}$$

Denote by $\frac{\mathbf{X}}{N}$ the probability function corresponding to the point in $\mathbb{D}^n$ defined by the vector

$$\left\langle \frac{X_1}{N}, \ldots, \frac{X_{2^n}}{N} \right\rangle$$

Now, taking the limit form of $\approx$ as $N \to \infty$ and $\epsilon \to 0$ we get

$$\sum_{j=1}^{2^k} x_{i,j} z_j = c_i \qquad (4.14)$$

$$\sum_{j=1}^{2^k} y_{i,j} z_j = 0 \qquad (4.15)$$

$$\sum_{j=1}^{2^k} z_j = 1 \qquad (4.16)$$

$$z_j \geq 0 \qquad (4.17)$$

Recall that in Definition 4.6 the inference process CFE_k is defined on constraint sets where each constraint has the form

$$\sum_{j=1}^{2^k} a_{i,j} w(\alpha_j) \leq b_i$$

There are also conditions which arise from w being a probability function. Namely,

$$\sum_{j=1}^{2^k} w(\alpha_j) = 1$$

$$w(\alpha_j) \geq 0$$

Hence the limit form of the expert constraint set E expressed by (4.14)–(4.17) defines a constraint set to which we can apply CFE$_k$ if we replace each equality with a pair of inequalities. For convenience let us now denote the set of constraints defined by (4.4)–(4.7) as Σ, the constraints defined by (4.10)–(4.13) as Σ', and the set defined by (4.14)–(4.17) as Σ''.

To develop our model further, we need to introduce the concept of an assignment of subsets corresponding to a c-frame. To that end we make the following definition

Definition 4.10 (C-distribution).

Let $E \in EX(N, n)$ and take a function $f : \mathrm{At}(E) \to \mathscr{P}(E)$. f is said to be a c-distribution if there is some c-frame $\overline{T}$ on L_n such that for each $i = 1, 2, \ldots, 2^n$

$$\alpha_i \in \mathrm{At}(L_n) \setminus \mathrm{At}(\overline{T}) \Rightarrow f(A_i) = \emptyset$$

where α_i runs through the atoms of L_k and A_i runs through the atomic properties of E. We assume an ordering of $\mathrm{At}(L_n)$ and $\mathrm{At}(E)$ to be such that for each $j = 1, 2, \ldots, n$

$$\alpha_i \models p_j \quad \Leftrightarrow \quad A_i \subseteq P_j$$

where $L_n = \{p_1, p_2, \ldots, p_n\}$ and the properties of E are $P_1, P_2, \ldots, P_n$. Denote the set of all c-distributions on E as $\mathfrak{CD}(E)$, and for an expert constraint set Σ denote the set of all c-distributions on E which are consistent with Σ by $\mathfrak{CD}(E, \Sigma)$.

The idea of course is that a c-distribution corresponding to $\overline{T}$ is an assignment of the examples in E in such a way that only assigns examples to the atoms of $\overline{T}$. This amounts to a constraint

$$\sum_{\alpha_j \in L_n \setminus \mathrm{At}(\overline{T})} z_j = 0 \tag{4.18}$$

Notice that this constraint is equally applicable to the expert constraints (4.10)–(4.13) and to the limit constraints in Σ''.

Each c-distribution also describes a probability function on L_k of course. We define the probability function $w_f \in \mathbb{D}^n$ by

$$w_f(\alpha_i) = \frac{|f(A_i)|}{N}$$

where the correspondence between α_i and A_i is as described above. Intuitively, a c-distribution corresponds to a c-frame iff its probability function is consistent with that c-frame.

Now, the intuitive idea behind this characterisation is that the probability functions derived from “most” c-distributions will be “close” to the solutions of Σ'' provided by CFE$_k$. However, it is not necessarily the case there will be a c-frame on L_k consistent with these constraints. To this end we need to enlarge the language until there is a c-frame which is consistent with Σ''. That is, we need to enlarge the language to the size of the molecular weight of Σ'', $\xi(\Sigma'')$ — this will exist iff Σ is a consistent set of constraints, due to a result in [35] which states that Σ'' will be consistent iff Σ is. Then there will be a c-distribution which is consistent with Σ.

Definition 4.11 (Expansions of constraint and example sets).

Take an example set $E \in EX(N,k)$. For $n > k$, the **expansion** of E to n properties is an example set $E_n \in EX(N,n)$ which has the same set of examples and properties $P_1, P_2, \ldots, P_k, P_{k+1}, \ldots, P_n$. The atomic properties of E_n correspond to those of E as follows. Suppose A is an atomic property of E. Then

$$A = \bigcup_{\substack{B \in \mathrm{At}(E_n),\\ B \subseteq A}} B$$

We can define an expansion of expert constraint sets directly from the above. For the expansion of $\Sigma \in EL(N,k)$ to n properties is defined as $\Sigma_n \in EL(N,n)$ where each constraint is defined from those in Σ by replacing each $A \in \mathrm{At}(E)$ by its expansion to $\mathrm{At}(E_n)$ as shown above. That is for every $A \in \mathrm{At}(E)$ each occurrence of $|A|$ in Σ is replaced in Σ_n by

$$\sum_{\substack{B \in \mathrm{At}(E_n),\\ B \subseteq A}} |B|$$

The relationship $\approx$ will also change for Σ_n — the condition (4.8) changes from

$$|X - Y| \leq 2^k \sqrt{m} \quad \Rightarrow \quad X \approx Y$$

to

$$|X - Y| \leq 2^n \sqrt{m} \quad \Rightarrow \quad X \approx Y$$

Finally, the expansion to n properties of the limit form of Σ is simply the limit form of the expansion to n properties of Σ.

We are now ready to state the theorem characterising CFE$_k$:

Theorem 4.1 (First Characterisation Theorem) *Assume* $\Sigma \in EL(N,k)$

is a consistent expert constraint set on an example set $E \in EX(N,k)$. Set $n = \xi(\Sigma'')$, where Σ'' is the limit form of Σ as described by (4.14)-(4.17), and let E_n,Σ_n and Σ''_n be the expansions to n properties of E, Σ and Σ'' respectively. Then for each $\mu, \nu > 0$ there exist N_0 and $\epsilon > 0$ such that for all $N \geq N_0$ and $\approx$ satisfying (4.8) and (4.9) there is a finite set of probability functions $\rho_1, \rho_2, \ldots, \rho_t \in \mathbb{D}^n$ for which the ratio

$$\frac{|\{f \in \mathfrak{CD}(E_n, \Sigma_n) \mid \|w_f - \rho_i\| \geq \nu \textit{ for all } \rho_i\}|}{|\mathfrak{CD}(E_n, \Sigma_n)|}$$

is at most μ.

Furthermore, the set of probability functions $\rho_1, \rho_2, \ldots, \rho_t$ is exactly the set chosen by $\mathrm{CFE}_n(\Sigma''_n)$.

Notice that since $n = \xi(\Sigma'')$, when restricted to L_k, each probability function ρ_i is a member of CFE_k. Hence this theorem does indeed characterise CFE_k as defined in Definition 4.6.

To begin the proof we first utilise Theorem 1 of [35]. For each c-frame $\overline{T}$ on L_n which is consistent with Σ''_n define $\Sigma_n^{\overline{T}}$ as being Σ_n together with the constraint:

$$\sum_{\alpha_j \in L_n \setminus \mathrm{At}(\overline{T})} |A_j| = 0 \tag{4.19}$$

Notice then that any $f \in \mathfrak{CD}(E_n, \Sigma_n^{\overline{T}})$ is consistent with $\overline{T}$. Suppose $\overline{\mathbf{T}}$ is the set of all c-frames on L_n which are consistent with Σ''_n. Now, applying Theorem 1 of [35] to $\Sigma_n^{\overline{T}}$ we get the result that

Lemma 4.2 *Suppose $\overline{T}$ is a c-frame on L_n which is consistent with Σ_n as above.Then for each $\mu, \nu > 0$ there exists $N_{\overline{T}}$ and $\epsilon_{\overline{T}}$ such that for all $N \geq N_{\overline{T}}$*

and $\approx$ satisfying (4.8) *and* (4.9) *the ratio*

$$\frac{\left|\left\{f \in \mathfrak{CD}(E_n, \Sigma_n^{\overline{T}}) \mid \|w_f - \rho_{\overline{T}}\| \geq \nu\right\}\right|}{\left|\mathfrak{CD}(E_n, \Sigma_n^{\overline{T}})\right|}$$

is at most $\mu/|\overline{\mathbf{T}}|$, where $\rho_{\overline{T}}$ is the maximum entropy solution on the limit form of $\Sigma_n^{\overline{T}}$.

Notice that the sets $\mathfrak{CD}(E_n, \Sigma_n^{\overline{T}})$, where $\overline{T}$ ranges over $\overline{\mathbf{T}}$, form a partition of $\mathfrak{CD}(E_n, \Sigma_n)$. This gives us

$$\begin{aligned}
&\frac{\left|\left\{f \in \mathfrak{CD}(E_n, \Sigma_n) \mid \|w_f - \rho_{\overline{T}}\| \geq \nu \text{ for every } \overline{T} \in \overline{\mathbf{T}}\right\}\right|}{|\mathfrak{CD}(E_n, \Sigma_n)|}\\
\leq &\frac{\sum_{\overline{T}\in\overline{\mathbf{T}}}\left|\left\{f \in \mathfrak{CD}(E_n, \Sigma_n^{\overline{T}}) \mid \|w_f - \rho_{\overline{T}}\| \geq \nu\right\}\right|}{\sum_{\overline{T}\in\overline{\mathbf{T}}}\left|\mathfrak{CD}(E_n, \Sigma_n^{\overline{T}})\right|}\\
\leq &\sum_{\overline{T}\in\overline{\mathbf{T}}}\frac{\left|\left\{f \in \mathfrak{CD}(E_n, \Sigma_n^{\overline{T}}) \mid \|w_f - \rho_{\overline{T}}\| \geq \nu\right\}\right|}{\left|\mathfrak{CD}(E_n, \Sigma_n^{\overline{T}})\right|}
\end{aligned}$$

where the last step is by Lemma A.1. By Lemma 4.2 if $N > \max\left\{N_{\overline{T}} \mid \overline{T} \in \overline{\mathbf{T}}\right\}$ and $\approx$ satisfies (4.8) and (4.9) for $\epsilon < \min\left\{\epsilon_{\overline{T}} \mid \overline{T} \in \overline{\mathbf{T}}\right\}$ then this is less than μ.

In other words, a weaker version of the First Characterisation Theorem follows immediately from Theorem 1 of [35] — namely that as N grows large, the c-distributions consistent with Σ_n will tend to cluster around the maximum entropy solutions of each c-frame consistent with Σ_n. However, we wish to prove something slightly stronger. Our characterisation theorem states that as N grows large, then certain of these c-frames will come to dominate, in the sense that more c-distributions will be near them. The rest of this section is the proof that the c-frames that do so dominate will be those selected by CFE$_k$. First we define a relationship that will be useful for this proof; that of a real number being "close to” a given integer.

Definition 4.12 (Closeness).

For the purposes of this proof, we define a relationship $\sim$ between the set of real numbers $\mathbb{R}$ and the integers $\mathbb{Z}$. For $x \in \mathbb{R}$ and $n \in \mathbb{Z}$, we say that “x is close to n”, denoted[7] $x \sim n$ if

$$n = [x] \qquad \text{or} \qquad n = [x] + 1$$

where $[x]$ denotes the smallest integer less than or equal to x.

We will require two Lemmas from the Proof of Theorem 1 in [35], namely Lemmas 2 and 3 on pages 21–23. We restate them here, slightly changing some notation:

Lemma 4.3 *Let $\mathbf{p} = (p_1, \ldots, p_{2^n})$, $0 \leq p_i \leq 1$ for all i, and $\Sigma_{i=1}^{2^n} p_i = 1$. Let $d = \min\{p_i \mid p_i \neq 0\}$ and $N \geq \frac{2}{d}$. Suppose $\mathbf{P} = (P_1, \ldots, P_{2^n})$ is a vector such that for all i, $P_i \sim Np_i$, $P_i = 0$ whenever $p_i = 0$ and $\sum_{i=1}^{2^n} P_i = N$. Then*

$$\left|\mathrm{h}_n(\mathbf{p}) - \mathrm{h}_n\left(\frac{\mathbf{P}}{N}\right)\right| \leq \frac{2^n}{N}\left|\ln\frac{d}{2}\right| + \frac{2}{dN}$$

Lemma 4.4 *Let $\mathbf{T} = (T_1, \ldots, T_{2^n})$ be a vector such that $\sum_{i=1}^{2^n} T_i = N$, where the T_i are non-negative integers. Let $\mathbf{t} = \frac{\mathbf{T}}{N}$. Then*

$$\left|\ln\begin{bmatrix} N \\ \mathbf{T}\end{bmatrix} - N.\,\mathrm{h}_n(\mathbf{t})\right| \leq \frac{2^n - 1}{2}\ln 2\pi N + \frac{2^n + 1}{4}$$

The following Lemma uses a method similar, yet not identical, to Lemma 4 of [35] to estimate how many frames cluster around the maximum entropy solutions

[7] We may sometimes write $n \sim x$ to mean the same thing. The meaning is apparent in any case - i.e., the real number is “close to” the integer.

corresponding to different c-frames.

Lemma 4.5 *Suppose that ρ is the maximum entropy solution corresponding to a c-frame $\overline{T_\rho}$ and σ is the same for $\overline{T_\sigma}$, where both $\overline{T_\rho}$ and $\overline{T_\sigma}$ are c-frames on L_n. Suppose further that there is some $\kappa > 0$ such that $\mathrm{h}_n(\rho) \geq \mathrm{h}_n(\sigma) + \kappa$.*

Now let $\mathbf{R}$ and $\mathbf{Q}$ be vectors of size 2^n where R_i, Q_i are non-negative integers and

$$\sum_{i=1}^{2^n} R_i = \sum_{i=1}^{2^n} Q_i = N$$

and suppose further that

$$R_i \sim N\rho_i \qquad \text{and} \qquad Q_i \sim N\sigma_i$$

Define d_R and d_Q to be

$$d_R = \min_{R_i > 0}\left(\frac{R_i}{N}\right) \qquad \text{and} \qquad d_Q = \min_{Q_i > 0}\left(\frac{Q_i}{N}\right)$$

Finally, if

$$N \geq \frac{2}{\kappa}\left\{(2^n - 1)\ln 2\pi N + 2^n\left|\ln\frac{d_R}{2}\right| + \frac{2}{d_R} + 2^n\left|\ln\frac{d_Q}{2}\right| + \frac{2}{d_Q}\right\}$$

Then

$$\ln\begin{bmatrix} N \\ \mathbf{R} \end{bmatrix} - \ln\begin{bmatrix} N \\ \mathbf{Q} \end{bmatrix} \geq \frac{1}{2}N\kappa$$

Proof. Define $f(\mathbf{R},\mathbf{Q}) := \ln\left[\begin{smallmatrix} N \\ \mathbf{R} \end{smallmatrix}\right] - \ln\left[\begin{smallmatrix} N \\ \mathbf{Q} \end{smallmatrix}\right]$. Then

$$\begin{aligned} f(\mathbf{R},\mathbf{Q}) = \ln\begin{bmatrix} N \\ \mathbf{R} \end{bmatrix} &- N.\,\mathrm{h}_n\left(\frac{\mathbf{R}}{N}\right) + N.\,\mathrm{h}_n\left(\frac{\mathbf{R}}{N}\right) \\ &- N.\,\mathrm{h}_n\left(\frac{\mathbf{Q}}{N}\right) + N.\,\mathrm{h}_n\left(\frac{\mathbf{Q}}{N}\right) - \ln\begin{bmatrix} N \\ \mathbf{Q} \end{bmatrix} \end{aligned}$$

From this re-arrangement we can see that

$$\begin{aligned}
f(\mathbf{R},\mathbf{Q}) \geq N.\left(\mathrm{h}_n\left(\frac{\mathbf{R}}{N}\right) - \mathrm{h}_n\left(\frac{\mathbf{Q}}{N}\right)\right) - \left|\ln\begin{bmatrix}N\\ \mathbf{R}\end{bmatrix} - N.\,\mathrm{h}_n\left(\frac{\mathbf{R}}{N}\right)\right| \\
- \left|\ln\begin{bmatrix}N\\ \mathbf{Q}\end{bmatrix} - N.\,\mathrm{h}_n\left(\frac{\mathbf{Q}}{N}\right)\right| \\
\geq N.\left(\mathrm{h}_n\left(\frac{\mathbf{R}}{N}\right) - \mathrm{h}_n\left(\frac{\mathbf{Q}}{N}\right)\right) - \left((2^n-1)\ln 2\pi N + \frac{2^n+1}{2}\right)
\end{aligned}$$

by Lemma 4.4. Now, we effect a further rearrangement of the terms in the first brackets to give us

$$\begin{aligned}
f(\mathbf{R},\mathbf{Q}) \geq N.\Bigg(\ \mathrm{h}_n\left(\frac{\mathbf{R}}{N}\right) - \mathrm{h}_n(\rho) + \mathrm{h}_n(\rho) \\
- \mathrm{h}_n(\sigma) + \mathrm{h}_n(\sigma) - \mathrm{h}_n\left(\frac{\mathbf{Q}}{N}\right)\Bigg) - \Delta
\end{aligned}$$

where

$$\Delta = (2^n-1)\ln 2\pi N + \frac{2^n+1}{2}$$

Hence

$$\begin{aligned}
f(\mathbf{R},\mathbf{Q}) &\geq N.\left(\mathrm{h}_n(\rho) - \mathrm{h}_n(\sigma)\right) \\
&- N.\left|\mathrm{h}_n(\rho) - \mathrm{h}_n\left(\frac{\mathbf{R}}{N}\right)\right| \\
&- N.\left|\mathrm{h}_n(\sigma) - \mathrm{h}_n\left(\frac{\mathbf{Q}}{N}\right)\right| - \Delta \\
&\geq N\kappa - \left(\Delta + 2^n\left|\ln\frac{d_R}{2}\right| + 2^n\left|\ln\frac{d_Q}{2}\right|\right)
\end{aligned}$$

by Lemma 4.3, and by definition of κ. Now, by our condition on the minimum size of N we see that

$$f(\mathbf{R},\mathbf{Q}) \geq \frac{1}{2}N\kappa$$

as required. □

Lemma 4.6 *Take $\Sigma \in CL_k$, and suppose $\delta, \nu > 0$. Let $\overline{T_\sigma}$ and $\overline{T_\rho}$ be c-frames on L_n, where $n \geq \xi(\Sigma)$, and let σ, ρ be the probability functions with maximum entropy in $V_n(\Sigma, \overline{T_\sigma})$ and $V_n(\Sigma, \overline{T_\rho})$ respectively. As in Lemma 4.5 suppose they are such that $\mathrm{h}_n(\rho) \geq \mathrm{h}_n(\sigma) + \kappa$ for some $\kappa > 0$.*

Now take some large integer N and let $\mathbf{X}$ range over all vectors of size 2^n such that X_i is a non-negative integer and $\sum_{i=1}^{2^n} X_i = N$. Then if $N \geq N_0$, where N_0 is such that

$$(2N_0\nu + 1)^{2^n} \exp\left(-\frac{1}{2} N_0 \kappa \right) \leq \delta \tag{$*$}$$

and

$$N_0 \geq \frac{2^n}{\nu^2} \tag{$\dagger$}$$

then

$$\frac{\sum \left\{ \left[{N \atop \mathbf{X}}\right] \;\middle|\; ||\frac{\mathbf{X}}{N} - \sigma|| < \nu,\ \frac{\mathbf{X}}{N} \in V(\Sigma, \overline{T_\sigma}) \right\}}{\sum \left\{ \left[{N \atop \mathbf{X}}\right] \;\middle|\; ||\frac{\mathbf{X}}{N} - \rho|| < \nu,\ \frac{\mathbf{X}}{N} \in V(\Sigma, \overline{T_\sigma}) \right\}} < \delta$$

Proof. First notice that for any $N \geq N_0$ the conditions $(*)$ and $(\dagger)$ still hold with N replacing N_0 throughout.

Now, there are less than $(2N\nu + 1)^{2^n}$ vectors $\mathbf{X}$ such that $||\frac{\mathbf{X}}{N} - \sigma|| < \nu$. Indeed, take some $\sigma_i \neq 0$. The range $(\sigma_i - N\nu, \sigma_i + N\nu)$ contains at most $2N\nu + 1$ integers. Hence there are at most $(2N\nu + 1)^{2^n}$ vectors $\mathbf{X}$ such that for each $1 \leq i \leq 2^n$, $X \in (\sigma_i - N\nu, \sigma_i + N\nu)$. These vectors will include at least all those that satisfy $||\frac{\mathbf{X}}{N} - \sigma|| < \nu$.

There is also at least one vector $\mathbf{X}$ such that $||\frac{\mathbf{X}}{N} - \rho|| < \nu$ since if a vector $\mathbf{X}$ is such that $X_i \sim N\rho_i$ then by $(\dagger)$

$$\left| \frac{Xi}{N} - \rho_i \right| \leq \frac{1}{N} \leq \frac{\nu^2}{2^n}$$

Clearly there is at least one vector $\mathbf{X}$ for which $X_i \sim N.\rho_i$ for each i, and so for such an $\mathbf{X}$ we see that

$$\left\| \frac{\mathbf{X}}{N} - \rho \right\| \leq \sqrt{\sum_{i=1}^{2^n} \frac{\nu^2}{2^n}} = \nu$$

Clearly by choosing N large enough we can pick such a vector to be arbitrarily close to $N\rho$ and hence $\frac{\mathbf{X}}{N}$ will be in $V(\Sigma, \overline{T_\rho})$ as this set is closed and convex. We choose one such vector and denote it $\mathbf{R}$.

Now consider the value $\begin{bmatrix} N \\ \mathbf{X} \end{bmatrix}$ for each of the vectors $\mathbf{X}$ in

$$\left\{ \mathbf{X} \;\middle|\; \left\| \frac{\mathbf{X}}{N} - \sigma \right\| < \nu,\ \frac{\mathbf{X}}{N} \in V(\Sigma, \overline{T_\sigma}) \right\}$$

Clearly, if $\mathbf{X}$ is such that $N\sigma_i \sim X_i$ for each i then Lemma 4.5 holds and we have

$$\begin{bmatrix} N \\ \mathbf{X} \end{bmatrix} \leq \begin{bmatrix} N \\ \mathbf{R} \end{bmatrix} \exp\left(-\frac{1}{2} N\kappa \right) \tag{‡}$$

Suppose however that $\mathbf{X}$ is such that $N\sigma_i \not\sim X_i$ for some i. There will be some probability function $\tau \in V(\Sigma, \overline{T_\sigma})$ for which $N\tau_i \sim X_i$ for every i. Moreover, since σ has maximum entropy on $V(\Sigma, \overline{T_\sigma})$ then there is some κ_1 for which $\mathrm{h}_n(\sigma) \geq \mathrm{h}_n(\tau) + \kappa_1$. Now take $\mathbf{S}$ to be a vector such that $N\sigma_i \sim S_i$ for each i. Then we can apply Lemma 4.5 again to $\mathbf{X}$ and $\mathbf{S}$ to get

$$\begin{bmatrix} N \\ \mathbf{X} \end{bmatrix} \leq \begin{bmatrix} N \\ \mathbf{S} \end{bmatrix} \exp\left(-\frac{1}{2} N\kappa_1 \right)$$

Together with (‡) this gives us

$$\begin{bmatrix} N \\ \mathbf{X} \end{bmatrix} \leq \begin{bmatrix} N \\ \mathbf{R} \end{bmatrix} \exp\left(-\frac{1}{2} N(\kappa + \kappa_1) \right)$$

and so we see that (‡) holds for all vectors $\mathbf{X}$ in

$$\left\{\mathbf{X}\ \middle|\ \left\|\frac{\mathbf{X}}{N}-\sigma\right\|<\nu,\ \frac{\mathbf{X}}{N}\in V(\Sigma,\overline{T_\sigma})\right\}$$

Hence the ratio

$$\frac{\sum\left\{\begin{bmatrix}N\\\mathbf{X}\end{bmatrix}\ \middle|\ \|\frac{\mathbf{X}}{N}-\sigma\|<\nu,\ \frac{\mathbf{X}}{N}\in V(\Sigma,\overline{T_\sigma})\right\}}{\sum\left\{\begin{bmatrix}N\\\mathbf{X}\end{bmatrix}\ \middle|\ \|\frac{\mathbf{X}}{N}-\rho\|<\nu,\ \frac{\mathbf{X}}{N}\in V(\Sigma,\overline{T_\sigma})\right\}}$$

is at most

$$\frac{(2N\nu+1)^{2n}\exp\left(-\frac{1}{2}N\kappa\right)\begin{bmatrix}N\\\mathbf{R}\end{bmatrix}}{\begin{bmatrix}N\\\mathbf{R}\end{bmatrix}}$$

which by condition $(*)$ is at most δ, as required. □

We are now ready to finish the proof of Theorem 4.1:

Proof of Theorem 4.1. Recall from the discussion of Lemma 4.2 that for each c-frame $\overline{T}$ on L_n which is consistent with Σ''_n there is a probability function $\rho_{\overline{T}}$ which has the maximum entropy on $V(\Sigma''_n,\overline{T})$. Let $\mathscr{T}$ denote the set of all c-frames on L_n which are consistent with Σ''_n. Then let $\mathscr{S}$ be that subset of $\mathscr{T}$ where the entropy of the probability function associated with each c-frame in $\mathscr{S}$ is the maximum possible. In other words, define

$$\mathscr{S}=\left\{\overline{T}\in\mathscr{T}\ \middle|\ \mathrm{H}_n(\rho_{\overline{T}})=\max_{\overline{T}\in\mathscr{T}}\mathrm{H}_n(\rho_{\overline{T}})\right\}$$

Then $\mathscr{S}$ is the set of c-frames whose corresponding probability functions form the set $\mathrm{CFE}_n(\Sigma''_n)$.

Now, we are interested in the ratio

$$\frac{\left|\left\{f\in\mathfrak{CD}(E_n,\Sigma_n)\ \middle|\ \|w_f-\rho_{\overline{T}}\|\geq\nu\text{ for all }\overline{T}\in\mathscr{S}\right\}\right|}{|\mathfrak{CD}(E_n,\Sigma_n)|}\tag{4.20}$$

To calculate (4.20) we count how many c-distributions have associated probability functions which are *close* to the $\rho_{\overline{T}}$'s first. Indeed, if we first set set, for each $\overline{S} \in \mathscr{T}$, $r_{\overline{S}}$ to be such that

$$\begin{aligned}&\left|\left\{f \in \mathfrak{C}\mathfrak{D}(E_n, \Sigma_n^{\overline{S}}) \;\middle|\; \|w_f - \rho_{\overline{T}}\| < \nu \text{ for some } \overline{T} \in \mathscr{T}\right\}\right| \\ &\qquad = r_{\overline{S}}\left|\left\{f \in \mathfrak{C}\mathfrak{D}(E_n, \Sigma_n^{\overline{S}}) \;\middle|\; \|w_f - \rho_{\overline{S}}\| < \nu\right\}\right| \qquad (4.21)\end{aligned}$$

then we can split the set of all c-distributions according to which c-frame they correspond to:

$$\begin{aligned}&\left|\left\{f \in \mathfrak{C}\mathfrak{D}(E_n, \Sigma_n) \;\middle|\; \|w_f - \rho_{\overline{T}}\| < \nu \text{ for some } \overline{T} \in \mathscr{T}\right\}\right| \\ &= \sum_{\overline{S}\in\mathscr{T}} r_{\overline{S}}\left|\left\{f \in \mathfrak{C}\mathfrak{D}(E_n, \Sigma_n^{\overline{S}}) \;\middle|\; \|w_f - \rho_{\overline{S}}\| < \nu\right\}\right| \\ &= \sum_{\overline{S}\in\mathscr{S}} r_{\overline{S}}\left|\left\{f \in \mathfrak{C}\mathfrak{D}(E_n, \Sigma_n^{\overline{S}}) \;\middle|\; \|w_f - \rho_{\overline{S}}\| < \nu\right\}\right| \\ &+ \sum_{\overline{S}\in\mathscr{T}\setminus\mathscr{S}} r_{\overline{S}}\left|\left\{f \in \mathfrak{C}\mathfrak{D}(E_n, \Sigma_n^{\overline{S}}) \;\middle|\; \|w_f - \rho_{\overline{S}}\| < \nu\right\}\right| \qquad (4.22)\end{aligned}$$

Now, as we discussed earlier, for each probability function

$$w_f = \langle w_1, w_2, \ldots, w_{2^n}\rangle$$

there are $\left[{N \atop \mathbf{F}}\right]$ distributions $f \in \mathfrak{C}\mathfrak{D}(E_n, \Sigma_n^{\overline{S}})$ s.t. $w_f = \frac{\mathbf{F}}{N}$, where

$$\mathbf{F} = \left\langle \frac{|f(A_1)|}{N}, \frac{|f(A_2)|}{N}, \ldots, \frac{|f(A_{2^n})|}{N}\right\rangle$$

where $A_1, A_2, \ldots, A_{2^n}$ are the atomic properties of E_n. Then, setting

$$\mathbb{X} = \left\{\mathbf{X} \in \mathbb{Z}^{2^n} \;\middle|\; \sum_{i=1}^{2^n} x_i = N,\ x_i \geq 0\right\}$$

we see that (4.22) becomes

$$
\begin{aligned}
&= \sum_{\overline{S}\in\mathscr{S}} r_{\overline{S}} \sum_{\mathbf{X}\in\mathbb{X}} \left| \left\{ \begin{bmatrix} N \\ \mathbf{X} \end{bmatrix} \;\middle|\; \left\| \frac{\mathbf{X}}{N} - \rho_{\overline{S}} \right\| < \nu,\ \frac{X}{N} \in V(\Sigma''_n, \overline{S}) \right\} \right| \\
&+ \sum_{\overline{S}\in\mathscr{T}\setminus\mathscr{S}} r_{\overline{S}} \sum_{\mathbf{X}\in\mathbb{X}} \left| \left\{ \begin{bmatrix} N \\ \mathbf{X} \end{bmatrix} \;\middle|\; \left\| \frac{\mathbf{X}}{N} - \rho_{\overline{S}} \right\| < \nu,\ \frac{X}{N} \in V(\Sigma''_n, \overline{S}) \right\} \right| \qquad (4.23)
\end{aligned}
$$

Since for each pair $\overline{S} \in \mathscr{S}$ and $\overline{T} \in \mathscr{S}$ it is the case that $\mathrm{H}_n(\rho_{\overline{S}}) > \mathrm{H}_n(\rho_{\overline{T}})$, then by Lemma 4.6 for any $\delta_{\overline{T}} > 0$ and sufficiently large N we have

$$
\frac{\sum \left\{ \left[\begin{smallmatrix} N \\ \mathbf{X} \end{smallmatrix}\right] \ \middle|\ ||\frac{\mathbf{X}}{N} - \rho_{\overline{T}}|| < \nu,\ \frac{\mathbf{X}}{N} \in V(\Sigma, \overline{T}) \right\}}{\sum \left\{ \left[\begin{smallmatrix} N \\ \mathbf{X} \end{smallmatrix}\right] \ \middle|\ ||\frac{\mathbf{X}}{N} - \rho_{\overline{S}}|| < \nu,\ \frac{\mathbf{X}}{N} \in V(\Sigma, \overline{S}) \right\}} < \delta_{\overline{T}}
$$

Let $\overline{R}$ be that c-frame in $\mathscr{S}$ for which

$$
r_{\overline{R}} \times \sum_{\mathbf{X}\in\mathbb{X}} \left\{ \begin{bmatrix} N \\ \mathbf{X} \end{bmatrix} \;\middle|\; ||\frac{\mathbf{X}}{N} - \rho_{\overline{R}}|| < \nu,\ \frac{\mathbf{X}}{N} \in V(\Sigma, \overline{R}) \right\}
$$

is minimal. Now, given $\delta > 0$, for each $\overline{T} \in \mathscr{T} \setminus \mathscr{S}$ choose $\delta_{\overline{T}}$ such that

$$
\delta_{\overline{T}} = \delta \frac{r_{\overline{R}}}{r_{\overline{T}}} . \frac{|\mathscr{S}|}{|\mathscr{T} \setminus \mathscr{S}|}
$$

Then for sufficiently large N we see that

$$\begin{aligned}
&\sum_{\overline{S}\in\mathscr{T}\setminus\mathscr{S}} r_{\overline{S}} \sum_{\mathbf{X}\in\mathbb{X}} \left|\left\{ \begin{bmatrix} N \\ \mathbf{X} \end{bmatrix} \;\middle|\; \left\| \frac{\mathbf{X}}{N} - \rho_{\overline{S}} \right\| < \nu,\ \frac{X}{N} \in V(\Sigma''_n, \overline{S}) \right\}\right| \\
&< \sum_{\overline{S}\in\mathscr{T}\setminus\mathscr{S}} r_{\overline{S}}\delta_{\overline{S}} \sum_{\mathbf{X}\in\mathbb{X}} \left|\left\{ \begin{bmatrix} N \\ \mathbf{X} \end{bmatrix} \;\middle|\; \left\| \frac{\mathbf{X}}{N} - \rho_{\overline{R}} \right\| < \nu,\ \frac{X}{N} \in V(\Sigma''_n, \overline{R}) \right\}\right| \\
&< \sum_{\overline{S}\in\mathscr{T}\setminus\mathscr{S}} r_{\overline{R}} \frac{|\mathscr{S}|}{|\mathscr{T}\setminus\mathscr{S}|}\delta \sum_{\mathbf{X}\in\mathbb{X}} \left|\left\{ \begin{bmatrix} N \\ \mathbf{X} \end{bmatrix} \;\middle|\; \left\| \frac{\mathbf{X}}{N} - \rho_{\overline{R}} \right\| < \nu,\ \frac{X}{N} \in V(\Sigma''_n, \overline{R}) \right\}\right| \\
&= \sum_{\overline{S}\in\mathscr{S}} r_{\overline{R}}\delta \sum_{\mathbf{X}\in\mathbb{X}} \left|\left\{ \begin{bmatrix} N \\ \mathbf{X} \end{bmatrix} \;\middle|\; \left\| \frac{\mathbf{X}}{N} - \rho_{\overline{R}} \right\| < \nu,\ \frac{X}{N} \in V(\Sigma''_n, \overline{R}) \right\}\right| \\
&< \delta \sum_{\overline{S}\in\mathscr{S}} r_{\overline{S}} \sum_{\mathbf{X}\in\mathbb{X}} \left|\left\{ \begin{bmatrix} N \\ \mathbf{X} \end{bmatrix} \;\middle|\; \left\| \frac{\mathbf{X}}{N} - \rho_{\overline{S}} \right\| < \nu,\ \frac{X}{N} \in V(\Sigma''_n, \overline{S}) \right\}\right|
\end{aligned}$$

Hence (4.23) is less than

$$\big(1+\delta\big) \sum_{\overline{S}\in\mathscr{S}} r_{\overline{S}} \sum_{\mathbf{X}\in\mathbb{X}} \left|\left\{ \begin{bmatrix} N \\ \mathbf{X} \end{bmatrix} \;\middle|\; \left\| \frac{\mathbf{X}}{N} - \rho_{\overline{S}} \right\| < \nu,\ \frac{X}{N} \in V(\Sigma''_n, \overline{S}) \right\}\right|$$

We now replace vectors $\mathbf{X}$ with c-distributions to see that this is equal to

$$\big(1+\delta\big) \sum_{\overline{S}\in\mathscr{S}} r_{\overline{S}} \left|\left\{ f \in \mathfrak{CD}(E_n, \Sigma_n^{\overline{S}}) \;\middle|\; \|w_f - \rho_{\overline{S}}\| < \nu \right\}\right| \tag{4.24}$$

Now, for each $\overline{S} \in \mathscr{S}$ set $t_{\overline{S}}$ and $s_{\overline{S}}$ to be such that

$$\begin{aligned}
&\left|\left\{ f \in \mathfrak{CD}(E_n, \Sigma_n^{\overline{S}}) \;\middle|\; \|w_f - \rho_{\overline{T}}\| < \nu \text{ for some } \overline{T} \in \mathscr{S} \right\}\right| \\
&\qquad\qquad = t_{\overline{S}} \left|\left\{ f \in \mathfrak{CD}(E_n, \Sigma_n^{\overline{S}}) \;\middle|\; \|w_f - \rho_{\overline{S}}\| < \nu \right\}\right|
\end{aligned}$$

and

$$\begin{aligned}&\Big|\Big\{f \in \mathfrak{C}\mathfrak{D}(E_n, \Sigma_n^{\overline{S}}) \ \big| \ \|w_f - \rho_{\overline{T}}\| < \nu \text{ for some } \overline{T} \in \mathscr{T} \setminus \mathscr{S}\Big\}\Big| \\ &\qquad\qquad\qquad = s_{\overline{S}} \Big|\Big\{f \in \mathfrak{C}\mathfrak{D}(E_n, \Sigma_n^{\overline{S}}) \ \big| \ \|w_f - \rho_{\overline{S}}\| < \nu\Big\}\Big|\end{aligned}$$

then $r_{\overline{S}} = t_{\overline{S}} + s_{\overline{S}}$. Of course, by Lemma 4.2, given any $\epsilon_{\overline{S}} > 0$ for sufficiently large N we have

$$s_{\overline{S}} < \epsilon_{\overline{S}}.t_{\overline{S}}$$

So (4.24) is less than

$$\begin{aligned}&\big(1+\delta\big) \sum_{\overline{S} \in \mathscr{S}} (1+\epsilon_{\overline{S}}) t_{\overline{S}} \Big|\Big\{f \in \mathfrak{C}\mathfrak{D}(E_n, \Sigma_n^{\overline{S}}) \ \big| \ \|w_f - \rho_{\overline{S}}\| < \nu\Big\}\Big| \\ =&\big(1+\delta\big) \sum_{\overline{S} \in \mathscr{S}} (1+\epsilon_{\overline{S}}) \Big|\Big\{f \in \mathfrak{C}\mathfrak{D}(E_n, \Sigma_n^{\overline{S}}) \ \big| \ \|w_f - \rho_{\overline{T}}\| < \nu \text{ for some } \overline{T} \in \mathscr{S}\Big\}\Big|\end{aligned}$$

Given some $\epsilon > 0$, then by appropriate choices of $\epsilon_{\overline{S}}$ this is then less than

$$\big(1+\delta\big)\big(1+\epsilon\big) \big|\big\{f \in \mathfrak{C}\mathfrak{D}(E_n, \Sigma_n) \ \big| \ \|w_f - \rho_{\overline{T}}\| < \nu \text{ for some } \overline{T} \in \mathscr{S}\big\}\big|$$

Hence, referring back to (4.22), we see that given $\epsilon, \delta > 0$, for sufficiently large N the following holds:

$$\begin{aligned}&\big|\big\{f \in \mathfrak{C}\mathfrak{D}(E_n, \Sigma_n) \ \big| \ \|w_f - \rho_{\overline{T}}\| < \nu \text{ for some } \overline{T} \in \mathscr{T}\big\}\big| \\ < \ &\big(1+\delta\big)\big(1+\epsilon\big) \big|\big\{f \in \mathfrak{C}\mathfrak{D}(E_n, \Sigma_n) \ \big| \ \|w_f - \rho_{\overline{T}}\| < \nu \text{ for some } \overline{T} \in \mathscr{S}\big\}\big|\end{aligned} \tag{4.25}$$

It will be more convenient for us to re-write (4.25) in the following way.

Given any $\Delta > 0$ then for sufficiently large N we have

$$\begin{aligned}
&\Big|\big\{f \in \mathfrak{CD}(E_n, \Sigma_n) \;\big|\; \|w_f - \rho_{\overline{T}}\| < \nu \text{ for some } \overline{T} \in \mathscr{S}\big\}\Big| \\
&> \big(1-\Delta\big)\Big|\big\{f \in \mathfrak{CD}(E_n, \Sigma_n) \;\big|\; \|w_f - \rho_{\overline{T}}\| < \nu \text{ for some } \overline{T} \in \mathscr{T}\big\}\Big| \qquad (4.26)
\end{aligned}$$

Returning now to (4.20), and setting $\Gamma = |\mathfrak{CD}(E_n, \Sigma_n)|$,we can now calculate that

$$\begin{aligned}
&\frac{\Big|\big\{f \in \mathfrak{CD}(E_n, \Sigma_n) \;\big|\; \|w_f - \rho_{\overline{T}}\| \geq \nu \text{ for all } \overline{T} \in \mathscr{S}\big\}\Big|}{\Gamma} \\
&= \frac{\Gamma - \Big|\big\{f \in \mathfrak{CD}(E_n, \Sigma_n) \;\big|\; \|w_f - \rho_{\overline{T}}\| < \nu \text{ for some } \overline{T} \in \mathscr{S}\big\}\Big|}{\Gamma} \\
&< \frac{\Gamma - \big(1-\Delta\big)\Big|\big\{f \in \mathfrak{CD}(E_n, \Sigma_n) \;\big|\; \|w_f - \rho_{\overline{T}}\| < \nu \text{ for some } \overline{T} \in \mathscr{T}\big\}\Big|}{\Gamma} \\
&= \frac{\Gamma - \big(1-\Delta\big)\Big|\big\{f \in \mathfrak{CD}(E_n, \Sigma_n) \;\big|\; \|w_f - \rho_{\overline{T}}\| < \nu \text{ for some } \overline{T} \in \mathscr{T}\big\}\Big|}{\Gamma}
\end{aligned}$$

$$\begin{aligned}
&= \frac{\Gamma - \big(1-\Delta\big)\Big[\Gamma - \Big|\big\{f \in \mathfrak{CD}(E_n, \Sigma_n) \;\big|\; \|w_f - \rho_{\overline{T}}\| \geq \nu \text{ for all } \overline{T} \in \mathscr{T}\big\}\Big|\Big]}{\Gamma} \\
&= \Delta + \big(1-\Delta\big)\frac{\Big|\big\{f \in \mathfrak{CD}(E_n, \Sigma_n) \;\big|\; \|w_f - \rho_{\overline{T}}\| \geq \nu \text{ for all } \overline{T} \in \mathscr{T}\big\}\Big|}{\Gamma}
\end{aligned}$$

Now, by the discussion after Lemma 4.2, we know that given any $\mu' > 0$, for sufficiently large N the fraction in the above equation is less than μ'. Hence (4.20) is less than

$$\Delta + (1-\Delta)\mu'$$

Setting $\mu' = \frac{\mu-\Delta}{1-\Delta}$ we then see that given $\mu, \nu > 0$, then for sufficiently large N

$$\frac{\Big|\big\{f \in \mathfrak{CD}(E_n, \Sigma_n) \;\big|\; \|w_f - \rho_{\overline{T}}\| \geq \nu \text{ for all } \overline{T} \in \mathscr{S}\big\}\Big|}{|\mathfrak{CD}(E_n, \Sigma_n)|} < \mu$$

This completes the proof of the First Characterisation Theorem. □

4.3.3 Analysis

We have provided here a proof that there is some “intuitive” justification for the adoption of CFE_k as an inference process. The First Characterisation Theorem states that as the size of our example sets grow without bound, an arbitrarily large proportion of them will produce probability functions arbitrarily close to those given by CFE_k on the same constraints. Our motivation for such a characterisation was discussed in terms of an “expert gambler” in Section 4.3.1. Theorem 4.1 re-phrased in this terminology essentially says that if our expert’s experience includes enough examples, then we should expect with probability arbitrarily close to 1 that that experience gives rise to a probability function arbitrarily close to a CFE_k solution of the knowledge embodied by the constraints given to us by our expert. In other words, under the assumption that the expert’s knowledge is structured as a c-frame and that they have “enough experience,” we should expect their conclusions to agree closely with (one of) the solutions provided by CFE_k.

While this seems to be an agreeable conclusion to reach, there are certain criticisms that can be made of this model of the reasoning process. While there are other important arguments to be made in criticism of this formulation, we save those until later. Instead we focus on one criticism here: namely, why do we only investigate the smallest possible c-frames, the Σ-minimal c-frame? It might be argued that to restrict our attention solely to those c-frames is simply an *ad hoc* constraint, and given the definition of CFE_k it is hardly surprising that the probability functions arising from this model correspond to these c-frames. Surely there might be larger c-frames which are more suited (in the sense of being more “common”) to a given c-frame?

In answer to this criticism, we could draw upon Ockham’s Razor, which states (see [29]):

Pluralitas non est ponenda sine neccesitate

which is translated as

“entities should not be multiplied unnecessarily”

In the context of this discussion, the fallacy of needless multiplication of entities can be invoked to ascribe a reason to restricting ourselves to the smallest c-frames. The smallest possible c-frames, the Σ-minimal c-frame, require the addition of the fewest possible propositional variables. That is, we have “multiplied entities” only up to the point where it becomes possible to apply our model of reasoning.

However, as we shall see in the next section, such a defence is unnecessary. If we accept the argument that it is unjustifiable to restrict our attention to the smallest c-frames, the natural approach now is to ask what happens when we consider *all* c-frames.

4.4 A Further Characterisation of CFE$_k$

We respond to the criticism that it is unjustified to restrict attention to the smallest c-frames here by giving another characterisation theorem for CFE$_k$. The concept here is that we allow the size of the c-frames under consideration to grow without bound. This allows us to consider assignments of the example sets which correspond to *all* possible c-frames, no matter their size. The theorem is as follows:

Theorem 4.7 (Second Characterisation Theorem) *Let $\Sigma \in EL(N,k)$ be a consistent expert constraint set on an example set $E \in EX(N,k)$, such that the limit form Σ'' of Σ is adamant. Set $n = \xi(\Sigma'')$ and take $M \gg n$. Let E_M,Σ_M*

and Σ''_M *be the expansions to* M *properties of* E, Σ *and* Σ'' *respectively. Then for each* $\mu, \nu > 0$ *there exist* N_0, M_0 *and* $\epsilon > 0$ *such that for all* $N \geq N_0$, $M \geq M_0$ *and* $\approx$ *satisfying* (4.8) *and* (4.9) *there is a finite set of probability functions* $\rho_1, \rho_2, \ldots, \rho_t \in \mathbb{D}^n$ *for which the ratio*

$$\frac{\left|\left\{f \in \mathfrak{CD}(E_M, \Sigma_M) \;\middle|\; \left\|w'_f - \rho_i\right\| \geq \nu \textit{ for all } \rho_i\right\}\right|}{|\mathfrak{CD}(E_M, \Sigma_M)|}$$

where $w'_f = w_f \upharpoonright L_n$*, is at most* μ*.*

Furthermore, the set of probability functions $\rho_1, \rho_2, \ldots, \rho_t$ *is exactly the set chosen by* $\mathrm{CFE}_n(\Sigma''_n)$*.*

In other words, if we insist upon considering *all* c-frames consistent with Σ of whatever size, then as we let the c-frames grow there comes a point where the majority of c-distributions consistent with the c-frames in question will determine probability functions arbitrarily close to those selected by CFE_k. Hence the problem raised by the criticism of the simpler characterisation of CFE_k described at the end of the last section is averted.

The outline of the proof is as follows: we consider certain c-frames on L_M which are certain type of extension of those of minimal size to L_M, which we call "natural extensions". We then show that these c-frames have the highest entropy of all c-frames on L_M: then by the arguments presented in the previous section, namely Lemma 4.6 and the proof of Theorem 4.1, the majority of c-distributions will be close to these natural extensions. Finally, we show that there are many more c-frames corresponding to extensions of the c-frames of minimal size which have highest entropy.

To begin the proof, we define what we mean by the extension of a c-frame to a larger language:

Definition 4.13 (Extension of a c-frame).

Suppose $\overline{T}$ is a c-frame on L_n. For $M > n$, $\overline{T^*}$ is an **extension** of $\overline{T}$ to L_M iff:

1. $\overline{T^*}$ is a c-frame on L_M;
2. $\overline{T^*} \models \overline{T}$
3. For each $q \in L_M \setminus L_n$, either:
   (a) $\overline{T^*} \models q \leftrightarrow p$ for some $p \in L_n$, or;
   (b) $\overline{T^*} \models q \to \alpha$ for some $\alpha \in \mathrm{At}(\overline{T})$.

The idea behind such a definition is of course that the extension of a c-frame $\overline{T}$ has the same structure as $\overline{T}$ on L_n — the extension simply adds further structure "below" the atoms of $\overline{T}$, or makes elements of L_M equivalent to L_n. In the course of the argument to follow we will have occasion to count the number of atoms of the extension which imply a given atom of the original c-frame. Hence, for an extension $\overline{T^*}$ of $\overline{T}$ we define, for each $\alpha \in \mathrm{At}(\overline{T})$

$$k_\alpha = \left|\left\{\beta \in \mathrm{At}(\overline{T^*}) \ \mid \beta \models \alpha\right\}\right|$$

Consider now a c-frame $\overline{T}$ on L_n and a probability function $w \in \mathbb{D}^n$ consistent with $\overline{T}$, and an extension $\overline{T^*}$ to L_M. With k_α defined as above define the function $w^* \in \mathbb{D}^M$ for each $\beta \in \mathrm{At}(\overline{T^*})$ as

$$w^*(\beta) = \frac{w(\alpha)}{k_\alpha}$$

where α is that atom of $\overline{T}$ such that $\beta \models \alpha$. Then by a simple maximum entropy argument w^* has the highest entropy of all probability functions in $\mathbb{D}^M$ which are consistent with $\overline{T^*}$ and whose restriction to L_n is w.

We now use a constructive argument to show that given a c-frame and a probability function on that c-frame, there is a certain type of extension whose corresponding probability functions have the highest entropy.

Lemma 4.8 *Let $\overline{T}$ be a c-frame on L_n and let $w \in V(\emptyset, \overline{T})$. Let $M \gg n$ and let $\overline{T_F}, \overline{T_G}$ be extensions of $\overline{T}$ to L_M. Define functions $F, G : \mathrm{At}(\overline{T}) \to \mathbb{N}$ by setting, for each $\alpha \in \mathrm{At}(\overline{T})$,*

$$F(\alpha) = |\left\{\beta \in \mathrm{At}(\overline{T_F}) \;\mid \beta \models \alpha\right\}|$$
$$G(\alpha) = |\left\{\beta \in \mathrm{At}(\overline{T_G}) \;\mid \beta \models \alpha\right\}|$$

Define probability functions w_F, w_G by the following:

$$w_F(\alpha) = \frac{w_\alpha}{F(\alpha)} \qquad \textit{for all } \alpha \in \mathrm{At}(\overline{T_F})$$
$$w_G(\alpha) = \frac{w_\alpha}{G(\alpha)} \qquad \textit{for all } \alpha \in \mathrm{At}(\overline{T_G})$$

Suppose the following also holds:

1. *There are distinct $\alpha, \beta \in \mathrm{At}(\overline{T})$ s.t.*
   - *(a) $F(\alpha) \geq w(\alpha).|\,\mathrm{At}(\overline{T_F})| + 1$*
   - *(b) $F(\beta) \leq w(\beta).|\,\mathrm{At}(\overline{T_F})| - 1$*
   - *(c) $G(\alpha) = F(\alpha) - 1$*
   - *(d) $G(\beta) = F(\beta) + 1$*
2. *For all $\gamma \in \mathrm{At}(\overline{T}) \setminus \{\alpha, \beta\}$, $F(\gamma) = G(\gamma)$.*

Then we have

$$\mathrm{H}_M(w_G) > \mathrm{H}_M(w_F)$$

Proof. For the purposes of this proof we will work with the log form of the entropy function h.

$$\begin{aligned}
\mathrm{h}_M(w_F) - \mathrm{h}_M(w_G) &= \sum_{\gamma \in \mathrm{At}(\overline{T_G})} w(\gamma) \ln w(\gamma) - \sum_{\gamma \in \mathrm{At}(\overline{T_F})} w(\gamma) \ln w(\gamma) \\
&= \sum_{\gamma \in \mathrm{At}(\overline{T})} w(\gamma) \big(\ln \frac{w(\gamma)}{G(\gamma)} - \ln \frac{w(\gamma)}{F(\gamma)}\big) \\
&= w(\alpha)(\ln \frac{w(\alpha)}{G(\alpha)} - \ln \frac{w(\alpha)}{F(\alpha)}) + w(\beta)(\ln \frac{w(\beta)}{G(\beta)} - \ln \frac{w(\beta)}{F(\beta)})
\end{aligned}$$

since $F(\gamma) = G(\gamma)$ for $\gamma \neq \alpha, \beta$. So

$$\begin{aligned}
\mathrm{h}_M(w_F) - \mathrm{h}_M(w_G) &= w(\alpha) \ln \frac{F(\alpha)}{G(\alpha)} + w(\beta) \ln \frac{F(\beta)}{G(\beta)} \\
&= w(\alpha) \ln \frac{F(\alpha)}{F(\alpha) - 1} + w(\beta) \ln \frac{F(\beta)}{F(\beta) + 1}
\end{aligned}$$

Now, by elementary calculus $\ln x \leq x - 1$, with equality iff $x = 1$. Hence

$$\begin{aligned}
\mathrm{h}_M(w_F) - \mathrm{h}_M(w_G) &< w(\alpha)\left(\frac{F(\alpha)}{F(\alpha) - 1} - 1\right) + w(\beta)\left(\frac{F(\beta)}{F(\beta) + 1} - 1\right) \\
&= \frac{w(\alpha)}{F(\alpha) - 1} - \frac{w(\beta)}{F(\beta) + 1} \\
&\leq \frac{1}{|\,\mathrm{At}(\overline{T_F})|} - \frac{1}{|\,\mathrm{At}(\overline{T_F})|} = 0
\end{aligned}$$

□

Corollary *The preceding Lemma clearly shows that given any extension $\overline{T_F}$ to L_M of a c-frame $\overline{T}$ as above, there is some extension $\overline{T^*}$ such that:*

1. *$w(\alpha).|\operatorname{At}(\overline{T^*})| \sim k_\alpha$, where*

   $$k_\alpha = |\left\{\beta \in \operatorname{At}(\overline{T^*}) \ \mid \beta \models \alpha\right\}|$$

   , and;

2. *With w_F defined as in the Lemma, and with w^* defined as*

   $$w^*(\beta) = \frac{w(\alpha)}{k_\alpha}$$

   for all $\beta \in \operatorname{At}(\overline{T^*})$ such that $\beta \models \alpha$, then we have

   $$\mathrm{H}_M(w^*) > \mathrm{H}_M(w_F)$$

The proof of this statement is by repeated application of the Lemma.

We now take the intuition afforded by this Corollary and define a set of extensions of a given c-frame which are "natural" with respect to a given probability function, in the sense that the probability functions associated with them have the highest entropy.

Definition 4.14 (Natural Extension of a c-frame).
Suppose $\overline{T}$ is a c-frame on L_n, and $w \in \mathbb{D}^n$ is a probability function consistent with $\overline{T}$. Now take $\overline{T^*}$ to be an extension of $\overline{T}$ to $L_M \supseteq L_n$. Then $\overline{T^*}$ is called a natural extension of $\overline{T}$ with respect to w if:

1. For every $\alpha \in \operatorname{At}(\overline{T^*})$ we have

   $$w(\alpha).|\operatorname{At}(\overline{T^*})| \sim k_\alpha$$

   where k_α is as defined in Definition 4.13, and;

2. For every $p \in L_M \setminus L_n$, p is an atom of $\overline{T^*}$

3. For every distinct pair $p, q \in L_k \setminus L_n$ $T* \models p \not\leftrightarrow q$.

Lemma 4.9 *Take $\Sigma \in CL_k$ adamant and $n \gg \xi(\Sigma)$. Then of those $w \in V_n(\Sigma)$ which correspond to c-frames of size n, $\mathrm{H}_n(w)$ will be greatest for those w whose corresponding c-frame is a natural extension of a Σ-minimal c-frame.*

Proof. Suppose $w \in V_n$ corresponds to some c-frame $\overline{T^*}$ on L_n. Then

$$\begin{aligned}\mathrm{H}_n(w) &= \prod_{\beta \in \mathrm{At}(T^*)} w(\beta)^{-w(\beta)} \\ &= \prod_{\alpha \in \mathrm{At}(\overline{T})} \prod_{\substack{\beta \in \mathrm{At}(\overline{T^*}),\\ \beta \models \alpha}} w(\beta)^{-w(\beta)}\end{aligned}$$

where $\overline{T}$ is the smallest c-frame s.t. $\overline{T^*}$ is an extension of $\overline{T}$ to L_n. Suppose $|\overline{T}| = t$.

Now suppose that $w(\alpha)$ is fixed for each $\alpha \in \mathrm{At}(\overline{T})$. Then $\mathrm{H}_n(w)$ takes it highest value when for each $\beta \in \mathrm{At}(\overline{T^*})$ we have

$$w(\beta) = \frac{w(\alpha)}{k_\alpha}$$

where $k_\alpha = |\left\{\beta \in \mathrm{At}(\overline{T^*}) \;\mid \beta \models \alpha\right\}|$. Then

$$\mathrm{H}_n(w) = \prod_{\alpha \in \mathrm{At}(\overline{T})} \left(\frac{w(\alpha)}{k_\alpha}\right)^{-w(\alpha)} = \mathrm{H}_t(w) \times \prod_{\alpha \in \mathrm{At}(\overline{T})} k_\alpha^{w(\alpha)}$$

Subject to the condition $\sum k_\alpha = |\,\mathrm{At}(\overline{T^*})|$, by the Corollary to Lemma 4.8

this is greatest when

$$k_\alpha \sim w(\alpha).|\operatorname{At}(\overline{T^*})|$$

Now set δ_α to be

$$\delta_\alpha = \frac{k_\alpha}{w(\alpha).|\operatorname{At}(\overline{T^*})|}$$

By the definition of $\sim$ in Definition 4.12 it is easy to see that

$$\delta_\alpha \in \left(1 - \frac{1}{w(\alpha).|\operatorname{At}(\overline{T^*})|}, 1 + \frac{1}{w(\alpha).|\operatorname{At}(\overline{T^*})|}\right)$$

Hence

$$\begin{aligned}\mathrm{H}_n(w) &= \mathrm{H}_t(w).|\operatorname{At}(\overline{T^*})|.\prod_{\alpha\in\operatorname{At}(\overline{T})}(\delta_\alpha.w(\alpha))^{w(\alpha)}\\ &= |\operatorname{At}(\overline{T^*})|.\prod_{\alpha\in\operatorname{At}(\overline{T})}\delta_\alpha^{w(\alpha)}\end{aligned}$$

Clearly, as $|\operatorname{At}(\overline{T^*})| \to \infty$ then $\delta_\alpha \to 1$, and so $\mathrm{H}_n(w) \to |\operatorname{At}(\overline{T^*})|$. Obviously, $|\operatorname{At}(\overline{T^*})| \leq n - t$, and so we require a sufficiently large n for this convergence to occur. Given such a n, $\mathrm{H}_n(w)$ is largest when $|\operatorname{At}(\overline{T^*})|$ is. This clearly occurs when $\overline{T^*}$ is a natural extension[8] of the smallest possible c-frame — that is, a Σ-minimal c-frame. □

The final part of the argument that we require here is to investigate how the number of natural extensions of a given c-frame and probability function grows with the size of the extensions. We will show in Lemma 4.10 that this number depends crucially on the entropy of the selected probability function. We make the following definition to allow us to count natural extensions which are "the

[8] Notice that it does not matter *which probability function* $\overline{T^*}$ is a natural extension with respect to — as their size grows they will all tend to have the same entropy.

same" with respect to a given c-frame.

Definition 4.15 ($\overline{T}$-isomorphism).

Take a c-frame $\overline{T}$ on L_n and let $\overline{T_1}$ and $\overline{T_2}$ be extensions of $\overline{T}$ to L_M for some $M > n$. A bijection $f : L_M \to L_M$ is called a $\overline{\mathbf{T}}$**-isomorphism** between $\overline{T_1}$ and $\overline{T_2}$ if it is constant on L_n and it preserves the structure of $\overline{T_1}$. More precisely, f is such that

$$f(p) = p \quad \text{for any } p \in L_n$$

and for any $\theta \in SL_M$

$$\overline{T_1} \models \theta \Leftrightarrow \overline{T_2} \models f(\theta)$$

where the extension of f to SL_M is defined in the usual inductive way, i.e. for $\theta, \phi \in SL_M$

- $f(\theta) = f(p)$ if $\theta = p \in L_M$
- $f(\neg\theta) = \neg f(\theta)$
- $f(\theta \wedge \phi) = f(\theta) \wedge f(\phi)$

Lemma 4.10 *Let $\overline{T}$ be a c-frame of size N and take $w_1, w_2 \in \mathbb{D}^n$ to be probability functions consistent with $\overline{T}$ s.t.*

$$\mathrm{H}_N(w_1) > \mathrm{H}_N(w_2)$$

Now take $M \gg N$. Let $\overline{T_1}, \overline{T_2}$ be natural extensions to L_M of w_1 and w_2 respectively. Let X_i denote the number of c-frames which are $\overline{T}$-isomorphic to $\overline{T_i}$. Then $X_1 > X_2$. Furthermore, as $M \to \infty$, $\frac{X_2}{X_1} \to 0$.

Proof.

$$X_i = \frac{|\,\mathrm{At}(\overline{T_i})|!}{\prod\limits_{\alpha \in \mathrm{At}(\overline{T})} k_\alpha^i!}$$

where $k_\alpha^i = |\left\{\beta \in \mathrm{At}(\overline{T_i}) \;\mid \beta \models \alpha\right\}|$ for all $\alpha \in \mathrm{At}(\overline{T})$.

As $\overline{T_1}$ and $\overline{T_2}$ are chosen so as maximise the entropy of their corresponding probability functions then

$$k_\alpha^i \sim w_i(\alpha).|\,\mathrm{At}(\overline{T_i})|$$

Hence by the Wallis Derivation (See Chapter 11 of [24]), as $|\,\mathrm{At}(\overline{T_i})|$ grows without bound then

$$X_i \to \Big(\mathrm{H}_N(w_i)\Big)^{|\,\mathrm{At}(\overline{T_i})|}$$

Now since $\overline{T_1}$ and $\overline{T_2}$ are natural extensions of the same c-frame $\overline{T}$ then $|\,\mathrm{At}(\overline{T_1})| = |\,\mathrm{At}(\overline{T_2})| = M - |\,\mathrm{At}(\overline{T})|$. Hence, as $M \to \infty$ we do indeed have $X_1 > X_2$. And of course we also see that as $M \to \infty$ then

$$\frac{X_2}{X_1} \to \left(\frac{\mathrm{H}_N(w_2)}{\mathrm{H}_N(w_1)}\right)^M \to 0$$

□

We are now ready to give a proof of Theorem 4.7. We give a heuristic proof using the previous two Lemmas and Theorem 4.1 to outline the argument. The majority of the work is done by the proof of Theorem 4.1 — here we characterise which are the c-frames of maximum entropy as n grows.

Proof of Theorem 4.7. Consider the c-distributions onto L_n for a constraint set Σ, where Σ'' is adamant and $n \geq \xi(\Sigma'')$. By Theorem 4.1, the majority of the c-distributions will be arbitrarily close to the CFE_n solutions of Σ''.

Now, Lemma 4.9 shows that for sufficiently large n these CFE_n solutions of Σ will correspond to c-frames which are natural extensions of the $Sigma''$-minimal c-frames. Hence, the majority of c-distributions will cluster around natural extensions of the Σ-minimal c-frames.

Finally then, we count which natural extensions of the Σ''-minimal c-frames are most common. Lemma 4.10 shows that as n grows the natural extensions which correspond to the probability functions on $L_{\xi(\Sigma'')}$ of highest entropy will come to dominate. Therefore, for sufficiently large n, then an arbitrarily high proportion of the c-distributions consistent with Σ_n will be arbitrarily close to natural extensions of the Σ''-minimal c-frames. This proves the theorem. □

4.5 Discussion

We have given an example in this chapter of an inference process defined on conjunctively closed frames, and presented two characterisations of its use as an attempt at providing some justification for it. The first justification suffers from the defect of being too simple in conception, although technically quite difficult. The assumption that we should only consider c-distributions onto Σ-minimal c-frames is far too limiting, and therefore it is no surprise that this characterisation agrees with the CFE process.

However, the second characterisation is much more sound. In effect we consider *all* possible c-frames which are consistent with our constraints, and show that the Σ-minimal c-frames still dominate.

Unfortunately the second characterisation, and the CFE process itself, are only defined on adamant constraint sets. While the first characterisation of CFE will work equally well on ethereal constraint sets, we have already seen that this is not really an adequate justification for the adoption of CFE as an inference process.

Now, it could be argued that adamant constraint sets do have a coherent meaning in terms of our original motivation discussed in Chapter 2, specifically the principles discussed in Section 2.2.4. If we are to consider CFE as a model of the perceptual process then it seems reasonable to insist that we do in fact perceive *something* - to receive no perceptions at all would correspond to total sensory deprivation (or perhaps death!). In this case, it would seem reasonable to insist that our inference process, based as it is upon direct perceptions and observations of the world, should fail. For example, the extensive horse-racing example considered in Section 4.3.1 does not describe an adamant constraint set, and so this constraint set would not be susceptible to analysis by CFE[9].

However, the condition that our constraint set is adamant still seems a rather severe one. The condition that we must perceive *something* is surely not part of our agent's knowledge, but rather a constraint on the inference process itself. It would be interesting therefore to investigate how the second characterisation would behave on ethereal constraint sets.

Some preliminary investigations seem to indicate that, for ethereal constraint sets defined on L_n, the process will assign the maximum possible probability consistent with the constraints to

$$\neg p_1 \wedge \neg p_2 \wedge \cdots \wedge \neg p_n$$

, and then behave as CFE on the other atoms. Unfortunately, we have no results concerning this conjecture yet. Also, it seems likely that this process will fail one

[9] Note that this does not disqualify the horse-racing discussion as a useful example of expert knowledge — only that the simplified knowledge presented in Section 4.3.1 is inadequate for analysis by CFE. It still serves to illustrate the nature of this conception of expert knowledge.

of our desiderata for inference processes, that of language invariance, which as we will see in the next chapter *does* hold for CFE on adamant constraints.

Chapter 5

Properties of CFE

In this chapter we discuss some of the properties of CFE, and what they mean for the process.

5.1 Language Invariance

Roughly speaking, an inference process is called *language invariant* if the addition of extra propositional variables to the language of the constraint set, but without the addition of any information about these new variables, does not change the probabilities it assigns to the original.

In the terminology of Chapter 4, this can be formulated as the following:

> Take N_k to be an indexed inference process, so that N_n is an inference process on L_n. Suppose that Σ is a constraint set on L_n. Now take $m > n$ and let Σ_m be the expansion of Σ to $L_m \supset L_n$. Then the family N_k is said to be language invariant if for all n and m we have
>
> $$(w \restriction L_n) \in N_n(\Sigma)$$

for each $w \in N_m(\Sigma_m)$.

Language Invariance is an important property of inference processes. Surely the probability one ascribes to a property should not be changed by considering some additional properties, about which we know nothing? Fortunately, CFE is language invariant:

Theorem 5.1 *CFE is Language Invariant*

***Proof*.** We give a heuristic proof. First note that if Σ is adamant then there is some subset $\{p_{i_1}, p_{i_2}, \ldots, p_{i_t}\}$ of L_n for which for every $w \in V_n(\Sigma)$ we have

$$w(p_{i_1} \vee p_{i_2} \vee \cdots \vee p_{i_t}) = 1 \tag{5.1}$$

Now, since for every $w' \in solsm(\Sigma_m)$, we have $w' \upharpoonright L_n \in V_n(\Sigma)$ then

$$w'(p_{i_1} \vee p_{i_2} \vee \cdots \vee p_{i_t}) = 1$$

and so Σ_m is also adamant. Therefore it is meaningful to talk of CFE being language invariant.

Now, for all $n \leq m \leq \xi(()\Sigma)$ we have

$$\mathrm{CFE}_m(\Sigma_m) = \mathrm{CFE}_{\xi(\Sigma)}(\Sigma_{\xi(\Sigma)}) \upharpoonright L_m$$

And so

$$\begin{aligned}\mathrm{CFE}_m(\Sigma_m) \upharpoonright L_n =& (\mathrm{CFE}_{\xi(\Sigma)}(\Sigma_{\xi(\Sigma)})) \upharpoonright L_n \\ =& \,\mathrm{CFE}_{\xi(\Sigma_{\xi(\Sigma)})} \upharpoonright L_n \\ =& \,\mathrm{CFE}_n(\Sigma)\end{aligned}$$

Hence CFE is trivially language invariant in this case.

Now, suppose that $m > \xi(\Sigma)$. By the preceding discussion we need only show that

$$\mathrm{CFE}_m(\Sigma_m) \restriction L_{\xi(\Sigma)} = \mathrm{CFE}_{\xi(\Sigma)}(\Sigma_{\xi(\Sigma)})$$

Consider then the Σ_m-minimal c-frames. Since $m > \xi(\Sigma)$, we will have $\xi(\Sigma_m) = m$. Then by the arguments presented for the proof of Theorem 4.7, each Σ_m-minimal c-frame will be a natural extension of a Σ-minimal c-frame. Furthermore, the probability function picked by CFE_m on each of these natural extensions will have maximum possible entropy and so, when restricted to $L_\xi(\Sigma)$, it will be an element of $\mathrm{CFE}_{\xi(\Sigma)}(\Sigma_{\xi(\Sigma)})$. Finally, each $\Sigma_{\xi(\Sigma)}$-minimal c-frame will have some natural extension to L_m, and hence we will indeed have

$$\mathrm{CFE}_m(\Sigma_m) \restriction L_{\xi(\Sigma)} = \mathrm{CFE}_{\xi(\Sigma)}(\Sigma_{\xi(\Sigma)})$$

□

5.2 Continuity

There is an argument to be made (see e.g. [7], [30], [34]) for the claim that inference processes should be in some sense "continuous" - that is, that small changes in the information contained in the constraint sets should only lead to small changes in the conclusion reached by the inference process, whatever "small" may mean in this context. Indeed, some of the simpler metrics which can be applied to measure closeness of constraint sets turn out to not quite capture the correct notion. A convincing argument is given in [34] that the correct metric to use is the Blaschke metric $\triangle$ on the solution sets defined by the constraints, which is defined on convex subsets C, D of $\mathbb{D}^n$ by

$$\triangle(C,D) = \inf\{\delta \ \mid \forall\vec{x} \in C\ \exists\vec{y} \in D, |\vec{x}-\vec{y}| \leq \delta \ \&\ \forall\vec{y} \in D\ \exists\vec{x} \in C, |\vec{x}-\vec{y}| \leq \delta\}$$

where $|\vec{x}-\vec{y}|$ is just the Euclidean distance between $\vec{x}$ and $\vec{y}$. The Blaschke metric essentially finds the smallest δ s.t. every point in C is at most δ from some point in D and every point in D is at most δ from some point in C.

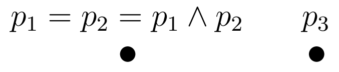
Then the continuity requirement as stated in [34] for an inference process N on a language L is

$$\textit{If } \theta \in SL, \quad K, K_i \in CL \textit{ for } i \in \mathbb{N} \quad \textit{and} \quad \lim_{i\to\infty} \triangle(V^L(K), V^L(K_i)) = 0$$

$$\textit{then} \quad \lim_{i\to\infty} N(K_i)(\theta) = N(K)(\theta)$$

However, as before, we should need to reformulate this definition to take into account that CFE may have multiple solutions. Such a reformulation is unnecessary though, as we can show immediately that CFE is not continuous in the above sense even when it has a single solution.

Let $\Sigma_\epsilon = \left\{w(p_1 \wedge p_2) \geq \epsilon, w(p_1) \geq \frac{1}{2} + \epsilon, w(p_2) \geq \frac{1}{2} + \epsilon\right\}$, where $0 < \epsilon < \frac{1}{2}$. Then Σ_ϵ is adamant and the molecular weight of Σ_ϵ is $\xi(\Sigma_\epsilon) = 3$, with the unique Σ_ϵ-minimal c-frame being $\overline{T_\epsilon}$ represented diagrammatically as



Now, also let $\Sigma = \left\{w(p_1) \geq \frac{1}{2}, w(p_2) \geq \frac{1}{2}\right\}$. Then Σ is adamant with $\xi(\Sigma) = 2$ and the unique Σ-minimal c-frame is $\overline{T}$, represented as

Hence we can see that

$$\begin{aligned} \mathrm{CFE}_2(\Sigma_\epsilon) &= \langle \frac{1}{2} + \epsilon, 0, 0, \frac{1}{2} - \epsilon \rangle \\ \mathrm{CFE}_2(\Sigma) &= \langle 0, \frac{1}{2}, \frac{1}{2}, 0 \rangle \end{aligned}$$

Now it is clear that the Σ_ϵ-minimal and Σ-minimal c-frames are very different. Indeed, $\overline{T_\epsilon}$ treats p_1 and p_2 as being equivalent, whereas $\overline{T}$ treats them as contradictory. However, the constraint sets that generate them are quite similar - it is easy to see that any solution of Σ_ϵ is a solution of Σ. Conversely, for any solution $\vec{x} \in \mathbb{D}^2$ of Σ, there is a solution $\vec{y}$ of Σ_ϵ s.t. $|x_i - y_i| \leq \epsilon$ for $i = 1, \ldots, 4$. Hence, in the Blaschke metric,

$$\triangle(V(\Sigma), V(\Sigma_\epsilon)) \leq 2\sqrt{\epsilon} \tag{5.2}$$

Now, let $\epsilon_i = \frac{1}{i+2}$ for $i = 1, 2, \ldots$. Then clearly

$$\lim_{i \to \infty} \triangle(V(\Sigma), V(\Sigma_{\epsilon_i})) \leq \lim_{i \to \infty} \frac{2}{\sqrt{i+2}} = 0$$

but

$$\begin{aligned} \lim_{i \to \infty} \mathrm{CFE}_2(\Sigma_{\epsilon_i}) &= \langle \frac{1}{2}, 0, 0, \frac{1}{2} \rangle \\ &\neq \langle 0, \frac{1}{2}, \frac{1}{2}, 0 \rangle = \mathrm{CFE}_2(\Sigma) \end{aligned}$$

Hence CFE does not satisfy this version of the continuity principle, in the case of CFE having a single solution.

5.3 Renaming

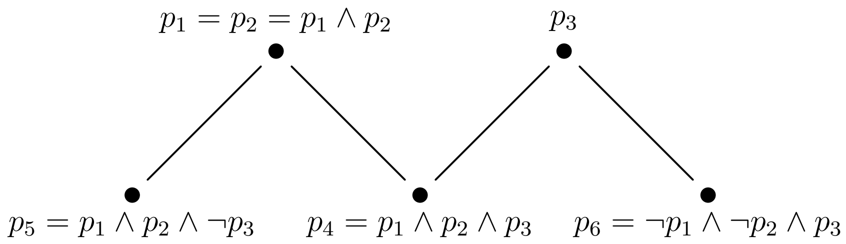
The renaming principle is stated in [34] for an inference process N on a language L as

> Suppose $K_1, K_2 \in CL$,
>
> $$K_1 = \left\{ \sum_{j=1}^{J} a_{ji} w(\gamma_j) = b_i \mid i = 1, \dots, m \right\},$$
> $$K_1 = \left\{ \sum_{j=1}^{J} a_{ji} w(\delta_j) = b_i \mid i = 1, \dots, m \right\},$$
>
> where $\gamma_1, \dots, \gamma_J, \delta_1, \dots, \delta_J$ are permutations of the atoms $\alpha_1, \dots, \alpha_J$ of L. Then
>
> $$N(K_1)(\gamma_j) = N(K_2)(\delta_j)$$

The justification given for this principle is that "the atoms of SL all share the same status of being simple possible worlds and so the particular ordering of $\alpha_1, \dots, \alpha_J$ of these atoms which we choose should not be significant. In a sense this principle can be viewed as a restricted version of the principle of indifference." In light of this, we should not expect CFE to satisfy this principle, as of course CFE is not indifferent to the atoms, and indeed some of the initial motivation for studying this inference process was that we should not be indifferent to renaming, especially with regard to negations. Now to show this by a concrete example, let Σ be given by

$$\{w(p_1 \wedge p_2 \wedge p_3) + w(p_1 \wedge p_2 \wedge \neg p_3) + w(\neg p_1 \wedge \neg p_2 \wedge p_3) = 1\}$$

Then Σ is adamant, and has molecular weight 6 with unique minimal c-frame $\overline{T}$ given by


$p_1 = p_2 = p_1 \wedge p_2$
p_3
$p_5 = p_1 \wedge p_2 \wedge \neg p_3$
$p_4 = p_1 \wedge p_2 \wedge p_3$
$p_6 = \neg p_1 \wedge \neg p_2 \wedge p_3$

Then, if $w = \mathrm{CFE}_3(\Sigma)$ we see that

$$w(p_1 \wedge p_2 \wedge p_3) = w(p_1 \wedge p_2 \wedge \neg p_3) = w(\neg p_1 \wedge \neg p_2 \wedge p_3) = \frac{1}{3}$$

and $w(\alpha) = 0$ for all other $\alpha \in \mathrm{At}(L_3)$.

However, consider

$$\Sigma' = \{w(p_1 \wedge p_2 \wedge p_3) + w(p_1 \wedge p_2 \wedge \neg p_3) + w(\neg p_1 \wedge \neg p_2 \neg \wedge p_3) = 1\}$$

obtained from Σ by exchanging the atoms $\neg p_1 \wedge \neg p_2 \wedge p_3$ and $\neg p_1 \wedge \neg p_2 \wedge \neg p_3$. Σ' is ephemeral, and so CFE gives only the solution

$$w(p_1) = w(p_2) = w(p_3) = 0$$
$$w(\neg p_1 \wedge \neg p_2 \neg \wedge p_3) = 1$$

and so CFE clearly does not satisfy this version of renaming. However, CFE may well satisfy some weaker versions of renaming. There are two particular versions of renaming that would seem appropriate to CFE. Firstly, what would happen if we were only to allow permutations of all the atoms *except* $\neg p_1 \wedge \neg p_2 \wedge \cdots \wedge \neg p_n$? It seems that this principle would clash strongly with the idea, embodied by the Classification Principle, that propositions and their negations are fundamentally different, and so CFE seems unlikely to satisfy it. Indeed, the following example shows just that.

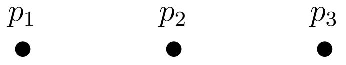
Let the constraint sets be defined as

$$\Sigma_1 = \left\{ w(p_1 \wedge p_2) + w(p_1 \wedge \neg p_2) = \frac{1}{2} \right\} \text{ and}$$
$$\Sigma_2 = \left\{ w(\neg p_1 \wedge p_2) + w(p_1 \wedge \neg p_2) = \frac{1}{2} \right\}$$

Then Σ_2 is clearly obtained from Σ_1 by the permutation of $\mathrm{At}(L_2)$ which only exchanges $p_1 \wedge p_2$ and $\neg p_1 \wedge p_2$.

Now, again $\xi(\Sigma_1) = 2$ and the unique Σ_1-minimal c-frame is

which has $\mathrm{CFE}_2(\Sigma_1) = \langle 0, \frac{1}{2}, \frac{1}{2}, 0 \rangle$. However, $\xi(\Sigma_2) = 3$ and the Σ_2-minimal c-frame is



and for this we have $\mathrm{CFE}(\Sigma_2) = \langle 0, \frac{1}{4}, \frac{1}{4}, \frac{1}{2} \rangle$, which is clearly not any permutation of the CFE solution to Σ_1, let alone the permutation which gives rise to Σ_2 from Σ_1.

Now that even this weakened principle of renaming has been shown to be too strong to be satisfied by CFE, we could consider the $n!$ permutations that only permute the p_i's - that is permutations of L_n, along with their natural extensions to SL_n. This principle would just amount to a relabelling of the elements of the Σ-minimal c-frames, and so should give the "same" solutions. However, a formal proof of this is yet to be produced.

Chapter 6

An Alternative Inference Process

In this chapter we present another inference process defined using the principles of maximum entropy and the concept of normal 1-frames, which we will designate as the **frame entropy** inference process. The motivation behind such an inference process is discussed and a technical definition of the process is given. A theorem is proved which characterises which normal 1-frames are the "most entropic."

Unfortunately, we will see that these "L-minimal" 1-frames, as we term them, are quite difficult to find — it is shown that there are only 3 such 1-frames in the case where L consists of only a single propositional variable, but we have found no characterisation of the L-minimal 1-frames for larger languages. However, a number of avenues of attack on this problem are presented along with some partial results which suggest that such a characterisation may be possible.

Finally we study how the inference process behaves in the case where L has only one variable. The process behaves quite unpredictably — we discuss the plausibility of eliciting any meaningful results from this inference process.

6.1 Concept

The inference process we will be investigating here will be a restriction of maximum entropy to the logical structures called normal 1-frames defined in Chapter 3. Due to the correspondence[1] between 1-frames as logical structures and simple undirected graphs with no isolated vertices, we shall tend to make no distinction between a normal 1-frame and its corresponding graph unless we need to.

In this chapter we will be interested primarily in *normal* 1-frames. For the sake of convenience it is to be understood that in this chapter when we refer to 1-frames we are actually referring to normal 1-frames.

The concept behind the inference process discussed here is very similar to that of the CFE inference process defined in Chapter 4. For CFE we restricted the maximum entropy process to consider only conjunctively closed positive frames. Here we consider a slightly different approach. The intuition is to restrict maximum entropy to normal 1-frames, capturing as they do the a simplified version of the Classification Principle, as discussed in Section 3.3. However, whereas in Chapter 4 we took the inference process as being defined on the smallest possible c-frames, in this chapter we shall look at an inference process which is defined on all possible 1-frames

6.2 Definitions and Notation

As in Chapter 4, we will be interested in probability functions that correspond to a certain type of logical structure — in this case normal 1-frames. If F is a 1-frame on a language L, and w is a probability function on a language $L' \supseteq L$

[1] See Corollary 3.8

then we say that **w is consistent with F** if for all $\alpha \in \mathrm{At}(L')$

$$w(\alpha) \neq 0 \Rightarrow \alpha \models F$$

That is, w only assigns non-zero weight to atoms which are consistent with F. For a constraint set Σ on L we defined the set of probability functions which satisfy Σ by $V^L(\Sigma)$, and where $L = L_n$ we write $V^n(\Sigma)$. Also, write $V^n(\Sigma, F)$ for the set of probability functions on L_n which satisfy Σ and are consistent with the normal 1-frame F. As an extension of this notation we will also write $\mathcal{V}^n(\Sigma)$ for the set of probability functions on L_m which satisfy Σ *and* which are consistent with some 1-frame on L_n.

Definition 6.1.

We define a two-dimensional series of inference processes. Define the $(\mathbf{n}, \mathbf{k})-$**frame entropy inference process** as $\mathrm{FE}_n^k : CL_n \to \mathscr{P}_0(\mathbb{D}^n)$ where

$$\mathrm{FE}_n^k(\Sigma) = \{w_1, w_2, \ldots, w_t\} \subseteq V^n(\Sigma)$$

s.t. each w_i is the restriction to L_n of some $w' \in \mathcal{V}^k(\Sigma)$ for which $\mathrm{H}_k(w')$ is maximal (i.e. there is no $w^* \in \mathcal{V}^k(\Sigma)$ for which $\mathrm{H}_k(w^*) > \mathrm{H}_k(w')$).

Let FE_n be the (infinite) sequence of inference processes

$$\left\langle \mathrm{FE}_n^n, \mathrm{FE}_n^{n+1}, \mathrm{FE}_n^{n+2}, \ldots \right\rangle$$

The motivation behind this definition is that we will examine how the terms of the infinite sequence FE_n behaves as $k \to \infty$. To examine which 1-frames will have the corresponding probability functions of maximum entropy, we now define an ordering on them. It will later turn out that the number of maximal independent sets that the graph of a given 1-frame has is crucial in determining which 1-frames are the "most entropic."

6.3 Ordering of 1-frames

Definition 6.2 (Weight of a valuation).

For a 1-frame G containing a language L and a valuation $v \in \mathrm{Val}\,(L)$, define

$$\mathrm{Val}\,(G)_v = \{w \in \mathrm{Val}\,(G) \ \mid w \restriction L = v\}$$

Define the weight of a valuation $v \in \mathrm{Val}\,(L)$ w.r.t. G to be

$$\|v\|_G = \left| \mathrm{Val}\,(G)_v \right|$$

For reasons of clarity we also define the following. Suppose G is a 1-frame which has a non-empty intersection with a language L. For a valuation $v \in \mathrm{Val}\,(L)$ we define the weight of v w.r.t. G as the weight w.r.t. G of v restricted to L:

$$\|v\|_G := \|v \restriction L\|_G$$

Definition 6.3 (L ordering).

We define an ordering $\prec_L$ on 1-frames containing a language L. For 1-frames F and G containing L set $F \preceq_L G$ iff $|F| \leq |G|$ and for all valuations $v \in \mathrm{Val}\,(L)$,

$$m\big(|G| - |F|\big).\ \|v\|_F \geq \|v\|_G$$

Finally, set $F \prec_L G$ iff $F \preceq_L G$ and $G \npreceq_L F$.

Recall from Section 3.3.2 that $m\,(n)$ is the Moon & Moser function for the maximum number of maximal independent sets on a graph of order n, and is given by $m\,(1) = 1$ and

$$m\,(n) = \begin{cases} 3^{\frac{n}{3}} & \text{if } n \equiv 0 \mod 3 \\ 4.3^{\frac{n-4}{3}} & \text{if } n \equiv 1 \mod 3 \\ 2.3^{\frac{n-2}{3}} & \text{if } n \equiv 2 \mod 3 \end{cases}$$

for $n \geq 2$. We also set $m\,(0) = 1$, since the empty graph has precisely one maximal independent set, namely $\emptyset$..

Definition 6.4 (L-minimal 1-frame).

A 1-frame F containing L is called L-minimal if there is no 1-frame $G \supset L$ s.t. $G \prec_L F$.

Definition 6.5 (Determination in G).

For a valuation $v \in \mathrm{Val}\,(L)$ and a 1-frame $G \supset L$, say v determines $q \in G \setminus L$ if every extension of v to G gives the same valuation to q. That is, for all $w_1, w_2 \in \mathrm{Val}\,(G)_v$, we have $w_1(q) = w_2(q)$.

Similarly, say v determines $Q \subseteq G \setminus L$ if v determines q for all $q \in Q$.

Lemma 6.1 *For any valuation $v \in \mathrm{Val}\,(L)$, if v determines $Q = \{q_1, \ldots, q_r\}$ in a 1-frame $G \supset L$, and there is no $q' \in G \setminus (L \cup Q)$ which is determined by v, then*

$$\|v\|_G \leq m\big(|G| - (|L| + r)\big)$$

Proof. Suppose $|G| = N$ and $|L| = n$. Every maximal independent set A of G is of the form $A = X \cup Y$, where $X = A \cap (L \cup Q)$ and $Y = A \setminus X$. Now, consider the set $\mathcal{V}$ of maximal independent sets of G which correspond to

v. That is,

$$\mathcal{V} = \{A \subset G \mid A \text{ is max. ind. in } G \text{ and } \forall p \in G\ v(p) = 1 \Rightarrow p \in A\}$$

Note that $|\mathcal{V}| = \|v\|_G$.

Now, for all $A_1, A_2 \in \mathcal{V}$, since v determines Q we see that $A_1 \cap (L \cup Q) = A_2 \cap (L \cup Q)$. So, if we set

$$\mathcal{V}' = \{A \setminus (L \cup Q) \mid A \in \mathcal{V}\}$$

it is clear that $|\mathcal{V}'| = |\mathcal{V}|$.

Let $G' = G \setminus (L \cup Q)$. Since every $A \in \mathcal{V}'$ is independent w.r.t. G it must also be independent w.r.t. G'.

Now, suppose A is not *maximal* independent w.r.t. G'. Then there is some $q' \in G'$ s.t. q' is not connected in G to any $p \in A$, and which *is* connected in G to some $p \in L \cup Q$ for which $w(p) = 1$ for all $w \in \text{Val}\,(G)_v$. Hence, $w(q') = 0$ for all $w \in \text{Val}\,(G)_v$: i.e., q' is determined by v, contradictory to assumption. So A is maximal independent w.r.t. G'.

Therefore $|\mathcal{V}'| \leq m\big(|G'|\big) = m\big(N - (n + r)\big)$. □

Proposition 6.2 *If $v \in \text{Val}\,(L)$ determines at least r elements of $G \setminus L$, then*

$$\|v\|_G \leq m\big(|G| - (|L| + r)\big)$$

Proof. Suppose the r elements of $G \setminus L$ determined by v are $\{q_1, \ldots, q_r\} = Q$. Let Q' be the set of *all* elements of $G \setminus L$ determined by v, and suppose $|Q'| = r'$.

Clearly $Q \subseteq Q'$ and $r \leq r'$. So, by Lemma 6.1,

$$\|v\|_G \leq m\big(|G| - (|L| + r')\big) \leq m\big(|G| - (|L| + r)\big) \qquad \square$$

6.4 A Characterisation of FE

In this section we present a Theorem which describes which 1-frames the FE process picks out as being "most entropic." Roughly speaking, we see that those 1-frames which have the "most efficient" distribution of their valuations will be the most entropic. First we make a definition:

Definition 6.6 (Connective Closure of a set in a 1-frame).
Suppose F is a normal 1-frame defined on L_k, and that X is some non-empty subset of L_k. The connective closure of X in F is that set which is "reachable" from some vertex of X. Formally, $y \in L_k$ is in the connective closure of X in F if there is some finite ordered set $\{z_1, z_2, \dots, z_t\} \subseteq L_k$ such that

1. $z_1 = y$ and $z_t \in X$

2. z_i and z_{i+1} are connected in the graph corresponding to F

We see that in the language of positive frames, the second condition of this definition becomes "z_{i+1} is a witness for z_i in F."

This definition allows us to capture the natural intuition that those elements of the language L_k which are "close" to the underlying language L_n in the incompatibility structure of F are more important than those which are not in determining which 1-frames are most entropic. We are now able to state our theorem.

Theorem 6.3 *Let Σ be a constraint set on L_n. The maximum frame entropy of a probability function on $L_k \supset L_n$ satisfying Σ occurs at a probability function w which is consistent with a 1-frame F of the form*

$$F \cong C \cup M_l$$

where C is the connective closure of L_n in F and is L_n-minimal, and M_l is the Moon-Moser graph of order $l = k - |C|$.

That is, $\mathrm{FE}_n^k(\Sigma)$ will be the restriction to L_n of a probability function on L_k consistent with such a frame.

Proof. Define $\mathrm{FH}_k(\Sigma)$ to be the maximum entropy of a probability function on L_k which satisfies Σ and is consistent with a 1-frame on L_k. We proceed as follows:

Let F be some fixed 1-frame on L_k. Define $\mathrm{FH}_k(\Sigma, F)$ to be the maximum entropy of a probability function on L_k which satisfies Σ and is consistent with F. Then

$$\mathrm{FH}_k(\Sigma) = \max\{\mathrm{FH}_k(\Sigma, F) \ \mid F \text{ is a 1-frame on } L_k\}$$

Now, by the definition of entropy,

$$\begin{aligned}
\mathrm{FH}_k(\Sigma, F) &= \max_{w \in V^k(\Sigma, F)} \left\{ \prod_{\beta \in \mathrm{At}(F)} w(\beta)^{-w(\beta)} \right\} \\
&= \max_{w \in V^k(\Sigma, F)} \left\{ \prod_{\substack{\alpha \in \mathrm{At}(L_n) \\ \not\models \neg(\alpha \wedge F)}} \prod_{\substack{\beta \in \mathrm{At}(F) \\ \beta \models \alpha}} w(\beta)^{-w(\beta)} \right\} \\
&= \max_{w \in V^n(\Sigma)} \left\{ \prod_{\substack{\alpha \in \mathrm{At}(L_n) \\ \not\models \neg(\alpha \wedge F)}} \max \left\{ \prod_{\substack{\beta \in \mathrm{At}(F) \\ \beta \models \alpha}} w'(\beta)^{-w'(\beta)} \;\middle|\; \begin{array}{c} w' \in V^n(\Sigma, F) \\ w' \restriction L_n = w \end{array} \right\} \right\} \\
&= \max_{w \in V^n(\Sigma)} \left\{ \prod_{\substack{\alpha \in \mathrm{At}(L_n) \\ \not\models \neg(\alpha \wedge F)}} \left(\frac{w(\alpha)}{\|v_\alpha\|_F} \right)^{-w(\alpha)} \right\}
\end{aligned}$$

where this last step is by an argument similar to that of the Corollary to Lemma 4.8.

Now, consider $G \succeq_n F$, $|G| = k$. Then $\|v_\alpha\|_G \leq \|v_\alpha\|_F$ for all $\alpha \in \mathrm{At}(L_n)$. Then, since any $w \in V_n(\Sigma)$ consistent with G is also consistent with F, we then have

$$F \preceq_n G \Rightarrow \mathrm{FH}_k(\Sigma, F) \geq \mathrm{FH}_k(\Sigma, G)$$

Also,

$$F \prec_n G \Rightarrow \mathrm{FH}_k(\Sigma, F) > \mathrm{FH}_k(\Sigma, G)$$

since there is some $\alpha \in \mathrm{At}(L_n)$ s.t. $\|v_\alpha\|_F > \|v_\alpha\|_G$. Hence,

$$\mathrm{FH}_k(\Sigma) = \max \left\{ \mathrm{FH}_k(\Sigma, F) \;\middle|\; \begin{array}{r} F \text{ is a 1-frame on } L_k \text{ for which} \\ \nexists \text{ a 1-frame } G \text{ on } L_k \text{ s.t. } G \prec_n F \end{array} \right\}$$

These 1-frames are exactly those of the form $F \cong C \cup M_l$. □

Now that we have shown that there is a characterisation of which 1-frames are important to the FE process, we examine how we may go about finding such frames. We begin with the one-dimensional case where Σ is defined on a language containing only one propositional variable.

6.5 The L_1-minimal frames

In this section, as we will be dealing exclusively with the language L_1 of order 1, for convenience we will write $\prec$ for $\prec_{L_1}$. First we note that the complete graphs of orders 2, 3 and 4 are L_1-minimal:

Proposition 6.4 *K_2, K_3 and K_4 are L_1-minimal, and are the only L_1-minimal frames of order $2, 3$ or 4.*

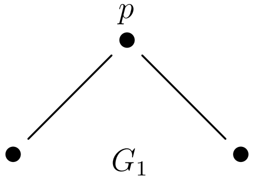
Proof. We take $L_1 = \{p\}$, and assume of course that the K_i's do actually contain L_1. For reasons of brevity write $\|1\|_F$ for the weight of v w.r.t. F when $v(p) = 1$. Similarly, write $\|0\|_F$ for the weight of v w.r.t. F when $v(p) = 0$.

1. **$\mathbf{K_2}$**

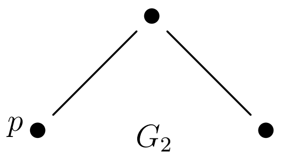
   It is clear that K_2 is L_1-minimal since there is no smaller 1-frame containing L_1, and K_2 is the only 1-frame of order 2. We have $\|1\|_{K_2} = \|0\|_{K_2} = 1$.

2. **$\mathbf{K_3}$**

   We have $\|1\|_{K_3} = 1$ and $\|0\|_{K_3} = 2$. Hence $K_2 \not\prec K_3$. Now, there are only two other 1-frames of order 3 which contain L_1, namely:



and



It is easy to see that for both G_1 and G_2 we have $\|1\|_{G_i} = \|0\|_{G_i} = 1$. Hence, for both G_i's we have $K_3 \prec G_i$, and so K_3 is indeed L_1-minimal, and is the only such graph of order 3.

3. $\mathbf{K_4}$

We have $\|1\|_{K_4} = 1$ and $\|0\|_{K_4} = 3$, and hence $K_2 \not\prec K_4$ and $K_3 \not\prec K_4$. Now, suppose there is some 1-frame G of order 4 containing L_1 s.t. $G \preceq_1 K_4$. Hence $\|1\|_G \geq 1$ and $\|0\|_G \geq 3$.

Now, $m(4) = 4$ and so in fact $\|1\|_G = 1$ and $\|0\|_G = 3$. Then we have $|\operatorname{Val}(G)| = 4 = m(4)$, and so G is the extremal Moon & Moser graph of order 4 - that is, $G = K_4$. So K_4 is indeed L_1-minimal.

Now take G to be any other graph G of order 4 which is L_1-minimal. Then $K_3 \not\prec$ and so either **a)** $\|1\|_G \geq 2$, or **b)** $\|0\|_G \geq 3$.

a) First, if $\|1\|_G = 3$, then $|\operatorname{Val}(G)| = 4 = m(4)$, and so G is the extremal Moon & Moser graph of order 4, K_4, contradictory to assumption. Hence $\|1\|_G = 2$. This gives $\|0\|_G = 1$, since otherwise we again get $G = K_4$. Now, $K_2 \prec G$, and so G is not L_1-minimal.

b) Obviously, $\|0\|_G \geq 3$ gives $|\operatorname{Val}(G)| = 4 = m(4)$, and so again $G = K_4$.

So K_4 is the only L_1-minimal graph of order 4. □

Next we see that any L_1-minimal 1-frame must correspond to a connected graph:

Proposition 6.5 *If $G \supset L_1 = \{p\}$ is L_1-minimal then G is connected.*

Proof. Notice that $G = H_1 \cup H_2$, where $p \in H_1$ and H_1 is the maximal connected subgraph of G which contains p — the connective closure of L_1 in fact.

Suppose G is disconnected: that is, $H_2 \neq \emptyset$. Then $H_1 \prec G$.

Indeed, for each $v \in \mathrm{Val}\,(L_1)$,

$$\|v\|_G = |\,\mathrm{Val}\,(H_2)\,|.\,\|v\|_{H_1}$$

and

$$\begin{aligned} &|\,\mathrm{Val}\,(H_2)\,| \leq m\big(|G| - |H_1|\big) \\ \Rightarrow &\|v\|_{H_1}.m\big(|G| - |H_1|\big) \geq \|v\|_G \\ \Rightarrow &H_1 \preceq G \end{aligned}$$

Since $H_1 \preceq G$ and $|H_1| < |G|$ then $H_1 \prec G$, and so G is not L_1-minimal. □

The last step is to show that there is a maximum size on L_1-minimal frames.

Theorem 6.6 *If $G \supset L_1 = \{p\}$ is a 1-frame of order $n \geq 5$ then G is not L_1-minimal.*

Proof. Suppose G *is* L_1-minimal. Then

$$\begin{aligned} K_2 \not\prec G \quad &\Rightarrow \quad \|0\|_G \geq m\,(n-2)+1 \textbf{ or } \|1\|_G \geq m\,(n-2)+1 &(6.1)\\ K_3 \not\prec G \quad &\Rightarrow \quad \|0\|_G \geq 2.m\,(n-3)+1 \textbf{ or } \|1\|_G \geq m\,(n-3)+1 &(6.2)\\ K_4 \not\prec G \quad &\Rightarrow \quad \|0\|_G \geq 3.m\,(n-4)+1 \textbf{ or } \|1\|_G \geq m\,(n-4)+1 &(6.3) \end{aligned}$$

Also, G is connected by Proposition 6.5 and so,

$$\|0\|_G + \|1\|_G \leq g\,(n) \qquad (6.4)$$

Now, the valuation $v(p) = 1$ determines at least one other element of G, and so $\|1\|_G \leq m\,(n-2)$ by Proposition 6.2. Hence, to satisfy (6.1) we must have

$$\|0\|_G \geq m\,(n-2)+1$$

Then (6.4) gives, by Lemma A.6,

$$\begin{aligned}\|1\|_G &\leq g\,(n) - (m\,(n-2)+1)\\ &< m\,(n-3)+1\end{aligned}$$

Hence to satisfy (6.2), we have

$$\|0\|_G \geq 2.m\,(n-3)+1$$

Again, (6.4) gives, by Lemma A.6,

$$\begin{aligned}\|1\|_G &\leq g\,(n) - (2.m\,(n-3)+1)\\ &< m\,(n-4)+1\end{aligned}$$

So to satisfy (6.3), we must now have

$$\begin{aligned}\|0\|_G &\geq 3.m\,(n-4)+1\\ &= m\,(n-1)+1 \text{ by Lemma A.4.}\end{aligned}$$

The valuation v which fixes $v(p)=0$ does not necessarily determine any elements of $G\backslash L_1$, and so by Proposition 6.2, we see that $\|0\|_G \leq m\,(n-1)$, which is a contradiction. Hence, G is not L_1-minimal. □

Theorem 6.7 *The only L_1-minimal 1-frames are K_2, K_3, and K_4.*

Proof. Immediate from Proposition 6.4 and Theorem 6.6. □

6.6 A Graph Theoretical approach to L-minimal frames

In this section, we will describe an attempt to construct L_n-minimal frames by means of successive graph operations.

Definition 6.7.

For a valuation $v \in \mathrm{Val}(F)$ of a 1-frame F and a maximal independent set X of F, we write $\mathbf{X} \models \mathbf{v}$ iff for all $p \in L_n$ we have $p \in X \Leftrightarrow v(p) = 1$. We say X **satisfies** v in this case.

Definition 6.8 (Moon-Moser Operation on a Graph).

Suppose F is a graph containing adjacent vertices x, y. Denote by $F(x; y)$ the graph obtained by:

1. Deleting all edges (x, z) where $z \neq y$, and;

2. Adding edges (x, z) for all $z \in N_F(y) \setminus \{x\}$.

This is called the Moon-Moser operation on F. It is fundamental to the argument presented in [32] where the graphs with the maximum number of maximal independent sets are determined.

Definition 6.9 (Depth of a vertex).

Let F be a graph defined on a set L_k, and let X be a non-empty proper subset of L_k. Suppose $p \in L_k \setminus X$. Then the depth of p from X is defined as being the minimum length of a path in F from p to any vertex in X, and we denote it $d(p, X)$.

If there is no such path, then we set $d(p, X)$ to be ∞.

The depth of a set of vertices X from a set of vertices Y is the minimal depth of any vertex $x \in X$ from Y, i.e.

$$d(X, Y) = \min_{x \in X} d(x, Y)$$

Our first result shows that we cannot destroy a valuation by removing vertices at maximal depth from L_n.

Proposition 6.8 *Suppose $F \supset L_n$ is a 1-frame and there is some $q \in F \setminus L_n$ for which $d(q, L_n)$ is maximal ≥ 2. Then, for any valuation $v \in \mathrm{Val}\,(L_n)$, if $\|v\|_F > 0$, then $\|v\|_{F \setminus q} > 0$.*

Proof. Take X to be a maximal independent set of F s.t. $X \models v$.

1. Suppose $q \notin X$. Then $X \setminus \{q\}$ is a maximal independent set of $F \setminus q$ which satisfies v. Hence $\|v\|_{F \setminus q} > 0$.
2. Suppose $q \in X$. Then $Y = X \setminus \{q\}$ is an independent set of $F \setminus q$. Suppose Y is not maximal such.

   Then there is $Z \subseteq N_F(q)$ s.t. Z is independent in F and no vertex in Z is adjacent to any in B. Pick a maximal such Z, and then $Y \cup Z$ is maximal independent in $F \setminus q$. Also, $d(q, L_n) \geq 2$ implies that $Z \cap L_n = \emptyset$. Hence $Y \cup Z \models v$, and so $\|v\|_{F \setminus q} > 0$. □

A leaf of a graph is a vertex of degree 1. The next result shows that if a 1-frame has a leaf at depth *geq*2 from L_n then conducting the Moon-Moser operation on this vertex produces a graph which precedes the original in the $\prec_n$ ordering.

Proposition 6.9 *Suppose a 1-frame $F \supset L_n$ has a leaf x at depth ≥ 2, and that x is connected to $y \in F \setminus L_n$. Then $F(x; y) \preceq_n F$.*

***Proof*.** Take $v \in Val(L_n)$. Let $G = F(x;y)$ and set

$$\begin{aligned} A_F =& \{X \text{ a max. ind. set of } F \mid X \models v,\ y \in X\} \\ B_F =& \{X \text{ a max. ind. set of } F \mid X \models v,\ y \notin X,\ N_F(y) \cap X = \{x\}\} \\ C_F =& \{X \text{ a max. ind. set of } F \mid X \models v,\ y \notin X,\ N_F(y) \cap X \setminus \{x\} \neq \emptyset\} \end{aligned}$$

Similarly, set

$$\begin{aligned} A_G =& \{X \text{ a max. ind. set of } G \mid X \models v,\ y \in X\} \\ B_G =& \{X \text{ a max. ind. set of } G \mid X \models v,\ x \in X\} \\ C_G =& \{X \text{ a max. ind. set of } G \mid X \models v,\ x, y \notin X\} \end{aligned}$$

Then,

$$\|v\|_F = |A_F| + |B_F| + |C_F| \tag{6.5}$$

$$\|v\|_G = |A_G| + |B_G| + |C_G| \tag{6.6}$$

and so

1. If $X \in A_F$ then $X \in A_G$, and vice-versa

$$\begin{aligned} &\Rightarrow A_F = A_G \\ &\Rightarrow |A_F| = |A_G| \end{aligned} \tag{6.7}$$

2. If $X \in B_F$ then $X \in B_G$, hence

$$|B_F| \leq |B_G| \tag{6.8}$$

3. If $X \in C_F$ then $X \setminus \{x\} \in C_G$, and so

$$|C_F| \leq |C_G| \tag{6.9}$$

Now, using equations (6.5) - (6.9), we see that

$$\begin{aligned} \|v\|_F &= |A_F| + |B_F| + |C_F| \\ &\leq |A_G| + |B_G| + |C_G| \\ &= \|v\|_G \end{aligned}$$

Hence, $G \preceq_n F$ as required. □

Now, a clique of a graph is a subset of its vertices, each of which is connected to every other vertex in the clique. A *maximal* clique is a clique which will cease to be a clique if we add any other vertex of the graph. The following result shows that certain types of clique can be removed from graphs to produce graphs which precede the original in the $\prec_n$ ordering:

Proposition 6.10 *Consider a 1-frame $F \supset L_n$ with vertex set $G \cup H$ which is the connective closure of L_n. Suppose that:*

1. *F has no leaves at depth ≥ 2;*
2. *H is a maximal clique of of F, and $r = |H| \geq 2$;*
3. *There is no maximal clique of F at greater depth than H, and the depth of H is ≥ 2, and;*
4. *There are exactly $1 \leq k < r$ edges (x_i, y_i) s.t. $x_i \in G$ and $y_i \in H$, and for $i \neq j$, $x_i \neq x_j$ and $y_i \neq y_j$.*

Then the 1-frame G_F induced by F on G has $G_F \prec_n F$.

***Proof*.** Pick $v \in \mathrm{Val}\,(L_n)$ s.t. $\|v\|_F > 0$. Now, let $\vec{\epsilon}$ range over $\{0,1\}^k$ and set:

$$\begin{aligned}
A_F^{\vec{\epsilon}} &= \{X \text{ max. ind. set of } F \mid X \models v,\ x_i \in X \Leftrightarrow \epsilon_i = 1\, y_1, \ldots, y_k \notin X\} \\
B_F &= \{X \text{ max. ind. set of } F \mid X \models v,\ x_1, \ldots, x_k, y_1, \ldots, y_k \notin X\} \\
C_F^{i,\vec{\epsilon}} &= \{X \text{ max. ind. set of } F \mid X \models v,\ y_i \in X, N_{G_F}(x_i) \cap X = \emptyset\} \\
D_F^{i,\vec{\epsilon}} &= \{X \text{ max. ind. set of } F \mid X \models v,\ y_i \in X, N_{G_F}(x_i) \cap X \neq \emptyset\}
\end{aligned}$$

Then

$$\|v\|_F = \sum_{\vec{\epsilon} \neq \vec{0}} |A_F^{\vec{\epsilon}}| + |B_F| + \sum_{i=1}^{k} \sum_{\vec{\epsilon}} \left(C_F^{i,\vec{\epsilon}} + D_F^{i,\vec{\epsilon}}\right) \tag{6.10}$$

Now also set

$$\begin{aligned}
A_G^{\vec{\epsilon}} &= \{X \text{ max. ind. set of } G \mid X \models v,\ x_i \in X \Leftrightarrow \epsilon_i = 1\} \\
B_G &= \{X \text{ max. ind. set of } G \mid X \models v,\ x_1, \ldots, x_k \notin X\}
\end{aligned}$$

Then

$$\|v\|_G = \sum_{\vec{\epsilon} \neq \vec{0}} |A_G^{\vec{\epsilon}}| + |B_G| \tag{6.11}$$

We proceed by restricting maximal independent sets of F to maximal independent sets of G, and counting them according to their type:

1. Take $X \in A_F^{\vec{\epsilon}}$. Then $X \cap G \in A_G^{\vec{\epsilon}}$. Further, for every $Y \in A_G^{\vec{\epsilon}}$, there are $r - k$ sets $X \in A_F^{\vec{\epsilon}}$ s.t. $X \cap G = Y$. Hence,

$$|A_F^{\vec{\epsilon}}| = (r-k)|A_G^{\vec{\epsilon}}| \tag{6.12}$$

2. Take $X \in B_F$. Then $X \cap G \in B_G$. Also, for every $Y \in B_G$, there are

$r-k$ sets $X \in B_F$ s.t. $X \cap G = Y$. Therefore,

$$|B_F| = (r-k)|B_G| \tag{6.13}$$

3. Take $X \in C_F^{i,\vec{\epsilon}}$.

A) Suppose $\epsilon_i = 1$. Then $X \supset \{x_i, y_i\}$, contradicting the independence of X. $\therefore$

$$|C_F^{i,\vec{\epsilon}}| = 0 \tag{6.14}$$

B) Suppose $\epsilon_i = 0$. Then $(X \cap G) \cup \{x_i\} \in A_G^{\vec{\epsilon}+\vec{e_i}}$, where $\vec{e_i}$ is the vector of length k with zeroes everywhere except in the ith position, which is 1. Hence,

$$\begin{aligned} |C_F^{i,\vec{\epsilon}}| &\leq |A_G^{\vec{\epsilon}+\vec{e_i}}| \\ \Rightarrow \sum_{\vec{\epsilon},\epsilon_i=0} |C_F^{i,\vec{\epsilon}}| &\leq \sum_{\vec{\epsilon},\epsilon_i=1} |A_G^{\vec{\epsilon}}| \end{aligned} \tag{6.15}$$

4. Take $X \in D_F^{i,\vec{\epsilon}}$.

A) Suppose $\epsilon_i = 1$. Then $X \supset \{x_i, y_i\}$, contradicting independence of X. So,

$$|D_F^{i,\vec{\epsilon}}| = 0 \tag{6.16}$$

B) Suppose $\epsilon_i = 0$ and $\vec{\epsilon} \neq \vec{0}$. Then $X \cap G \in A_G^{\vec{\epsilon}}$, and therefore

$$|D_F^{i,\vec{\epsilon}}| \leq |A_G^{\vec{\epsilon}}| \tag{6.17}$$

C) Suppose $\vec{\epsilon} = \vec{0}$. Then $X \cap G \in B_G$ and for all $Y \in B_G$ there is exactly one $X \in D_F^{i,\vec{\epsilon}}$ s.t. $X \cap G = Y$, namely $X = Y \cup \{y_i\}$.

Hence,

$$|D_F^{i,\vec{0}}| = |B_G| \tag{6.18}$$

Substituting the above into (6.10), we see that, by (6.12) and (6.13),

$$\|v\|_F = (r-k)\Bigg(|B_G| + \sum_{\vec{\epsilon}\neq\vec{0}} |A_G^{\vec{\epsilon}}|\Bigg) + \sum_{i=1}^{k}\sum_{\vec{\epsilon}}\Bigg(|C_F^{i,\vec{\epsilon}}| + |D_F^{i,\vec{\epsilon}}|\Bigg)$$

Rearranging, and using (6.14), (6.16), (6.17) and (6.18),

$$\begin{aligned}\|v\|_F \leq r|B_G| + \frac{r-k}{k}\sum_{i=1}^{k}\Bigg(\sum_{\substack{\vec{\epsilon}\neq\vec{0},\\ \epsilon_i=0}}|A_G^{\vec{\epsilon}}| + \sum_{\vec{\epsilon},\epsilon_i=1}|A_G^{\vec{\epsilon}}|\Bigg)\\ + \sum_{i=1}^{k}\Bigg[|C_F^{i,\vec{0}}| + \sum_{\vec{\epsilon}\neq\vec{0},\epsilon_i=0}\big(|C_F^{i,\vec{\epsilon}}| + |A_G^{\vec{\epsilon}}|\big)\Bigg]\end{aligned}$$

which gives

$$\begin{aligned}\|v\|_F \leq r|B_G| + \frac{r}{k}\Bigg(\sum_{i=1}^{k}\sum_{\substack{\vec{\epsilon}\neq\vec{0},\\ \epsilon_i=0}}|A_G^{\vec{\epsilon}}|\Bigg) + \frac{r-k}{k}\Bigg(\sum_{i=1}^{k}\sum_{\vec{\epsilon},\epsilon_i=1}|A_G^{\vec{\epsilon}}|\Bigg)\\ + \sum_{i=1}^{k}\sum_{\epsilon_i=0}|C_F^{i,\vec{\epsilon}}|\end{aligned}$$

and so by (6.15),

$$\begin{aligned}
\|v\|_F &\leq r|B_G| + \frac{r}{k}\sum_{i=1}^{k}\left(\sum_{\substack{\vec{\epsilon}\neq\vec{0},\\ \epsilon_i=0}}|A_G^{\vec{\epsilon}}| + \sum_{\vec{\epsilon},\epsilon_i=1}|A_G^{\vec{\epsilon}}|\right)\\
&= r|B_G| + \frac{r}{k}\sum_{i=1}^{k}\sum_{\vec{\epsilon}\neq\vec{0}}|A_G^{\vec{\epsilon}}|\\
&= r\left(|B_G| + \sum_{\vec{\epsilon}\neq\vec{0}}|A_G^{\vec{\epsilon}}|\right)\\
&= r\,\|v\|_{G_F}\\
&\leq m(r)\,\|v\|_{G_F}
\end{aligned}$$

Then, since also $|G| \leq |F|$, we have $G_F \prec_n F$ as claimed. □

Unfortunately, we have managed no further results in this vein. It may be possible to use this technique to give a constructive process for finding L_n-minimal graphs, but as yet we cannot see how. The results presented in this section lead us to conjecture that the maximal depth of vertex in an L-minimal frame from L is 1. A major step toward this conjecture this would be an answer to the following question:

Conjecture *If we remove maximal cliques at maximal depth from L, no matter what the connections between the clique and the rest of the graph, does the resulting graph precede the original in the $\prec_L$ ordering?*

6.7 Toward a Logical Characterisation of L_n-minimal frames

In this section, we turn away from the graph-theoretical characterisation of 1-minimal frames, and consider the logical properties that L-minimal frames must have. It was hoped that this would lead to a characterisation of L-minimal frames allowing us to find them.

Definition 6.10 (p-Negation).

For a normal 1-frame F, set N_F for each $p \in F$ to be the set of witnesses for $p \in F$.

Now, for any $p, q \in F$ let $\bigvee(N_F(q) \setminus \{p\})$ be denoted by $\mathrm{Neg}_p(q)$, called the "p-negation of q."

Note that if $p \notin N_F(q)$ then $\mathrm{Neg}_p(q) \equiv q$. Also, if $N_F(q) = \{p\}$ then $\mathrm{Neg}_p(q) \equiv \bot$.

First we state a result that allows us to make connections between the different orderings on 1-frames:

Lemma 6.11 *Suppose F, G are 1-frames with $L_{n+1} \subseteq |F| \subseteq |G|$. Then*

$$F \prec_{n+1} G \Rightarrow F \prec_n G$$

Proof. Relabel $|F|$ and $|G|$ so that $|F| = L_{n+k}$ and $|G| = L_{n+l}$ where $1 \leq k \leq l$.

Since $|F| \subseteq |G|$ and $F \prec_{n+1} G$, then

$$\forall v \in \mathrm{Val}\left(L_{n+1}\right),\ \left\|v\right\|_F .m\left(l-k\right) \geq \left\|v\right\|_G \tag{6.19}$$

$$\text{If } l = k \text{ then } \exists w \in \mathrm{Val}\left(L_{n+1}\right),\ \left\|w\right\|_F > \left\|w\right\|_G \tag{6.20}$$

Consider $\alpha \in \mathrm{At}(L_n)$. There are (unique) $\alpha^+, \alpha^- \in \mathrm{At}(L_{n+1})$ s.t. $\alpha \equiv \alpha^+ \vee \alpha^-$, namely $\alpha^+ = \alpha \wedge p_{n+1}$ and $\alpha^- = \alpha \wedge p_{n+1}$.

For any atom α of a language L, let v_α be the valuation in $\mathrm{Val}\left(L\right)$ corresponding to α in the usual way. Then it is easy to see that for any

$\alpha \in \mathrm{At}(L_n)$ and any 1-frame H s.t. $|H| \supseteq L_{n+1}$,

$$\|v_\alpha\|_H = \|v_{\alpha^+}\|_H + \|v_{\alpha^-}\|_H \tag{6.21}$$

Now, (6.19) and (6.21) give, for all $\alpha \in \mathrm{At}(L_n)$

$$\begin{aligned}\|v_\alpha\|_F .m\,(l-k) &= \|v_{\alpha^+}\|_F .m\,(l-k) + \|v_{\alpha^-}\|_F .m\,(l-k)\\ &\geq \|v_{\alpha^+}\|_G + \|v_{\alpha^-}\|_G\\ &= \|v_\alpha\|_G\end{aligned}$$

That is

$$\forall v \in \mathrm{Val}\,(L_n)\,,\ \|v\|_F .m\,(l-k) \geq \|v\|_G \tag{6.22}$$

So if $l > k$ then the claim is proved.

Suppose now $l = k$ and pick $w \in \mathrm{Val}\,(L_{n+1})$ s.t. $\|w\|_F > \|w\|_G$, which exists by (6.20). Then there is $\alpha \in \mathrm{At}(L_n)$ s.t. either $w = v_{\alpha^+}$ or $w = v_{\alpha^-}$. Suppose the former holds. Then by (6.19),(6.20) and (6.21)

$$\begin{aligned}\|v_\alpha\|_F &= \|w\|_F + \|v_{\alpha^-}\|_F\\ &> \|w\|_G + \|v_{\alpha^-}\|_G\\ &= \|v_\alpha\|_G\end{aligned}$$

This argument clearly also holds if $w = v_{\alpha^-}$. Hence if $l = k$

$$\exists v \in \mathrm{Val}\,(L_n)\,,\ \|v\|_F > \|v\|_G \tag{6.23}$$

namely $v = v_\alpha$. Considering (6.22) and (6.23) together the claim is also proved for the case $l = k$. □

By applying induction with the Lemma 6.11 we get the following immediate corollary.

Corollary 6.12 *Suppose F, G are 1-frames with $L_{n+1} \subseteq |F| \subseteq |G|$. Then*

$$F \prec_{n+k} G \Rightarrow F \prec_n G$$

The following Lemma describes how the valuations of a 1-frame must change if one of its incompatibilities is deleted:

Lemma 6.13 *Let F be a 1-frame with distinct elements $p, q \in |F|$ s.t. $p \in N_F(q)$. Suppose F' is obtained from F by deleting the edge $\{p, q\}$ and $R = \mathrm{Val}\,(F) \setminus \mathrm{Val}\,(F')$ is non-empty. Then for all $v \in R$*

$$\begin{aligned} &\textit{either} \quad v(p) = 1 \quad \textit{and} \quad v(q) = v(\mathrm{Neg}_p(q)) = 0 \\ &\textit{or} \quad v(q) = 1 \quad \textit{and} \quad v(p) = v(\mathrm{Neg}_q(p)) = 0 \end{aligned}$$

Proof. Consider $v \in R$. Let X be the maximal independent set of F corresponding to v (i.e. $X = \{p \in |F| \mid v(p) = 1\}$). Now since $v \notin R$, X is not maximal independent in F'. But since X is still independent in F' there must be $x \in |F| \setminus X$ s.t. $X \cup \{x\}$ is independent in G.

Let H be the graph formed from F by deleting the vertices p and q. Then for any set $A \subseteq |H|$ it is easy to see that the following are equivalent:

1. A is independent in H
2. A is independent in F
3. A is independent in G

Suppose now that $x \notin \{p, q\}$. Then, since $X \cup \{x\}$ is independent in G

$$\begin{aligned} & (X \cup \{x\}) \setminus \{p, q\} \text{ is independent in G} \\ \Rightarrow \quad & (X \cup \{x\}) \setminus \{p, q\} \text{ is independent in F} \end{aligned} \tag{6.24}$$

Now suppose $p \in X$. Then $x \notin N_G(p)$ since otherwise $X \cup \{x\}$ would not be independent in G. Similarly if $q \in X$ then $x \notin N_G(q)$. Then, since also $x \neq p, q$, (6.24) gives

$$X \cup \{x\} \text{ is independent in F}$$

which contradicts the maximality of X. Hence $x \in \{p, q\}$.

Suppose, wolog, that $x = p$. Then $p \notin X$

$$\Rightarrow \quad v(p) = 0 \tag{6.25}$$

Also, since $X \cup \{p\}$ is independent in G we have $N_G(p) \cap X = \emptyset$

$$\Rightarrow \quad v(\mathrm{Neg}_q(p)) = 0 \tag{6.26}$$

Next, since X is *maximal* independent in F and $p \notin X$, we must have

$$N_F(p) \cap X \neq \emptyset$$

But $N_F(p) = N_G(p) \cup \{q\}$. So the above gives

$$\begin{aligned} & \{q\} \cap X \neq \emptyset \\ \Rightarrow \quad & q \in X \\ \Rightarrow \quad & v(q) = 1 \end{aligned} \tag{6.27}$$

So (6.25),(6.26) and (6.27) give

$$v(q) = 1 \quad \text{and} \quad v(p) = v(\mathrm{Neg}_q(p)) = 0$$

as required. Similarly, if $x = q$ then

$$v(p) = 1 \quad \text{and} \quad v(q) = v(\mathrm{Neg}_p(q)) = 0 \qquad \square$$

We now present a series of lemmas stating various properties of L_n-minimal 1-frames discovered by considering various operations upon their graphs. These results are used later to give some limited characterisation theorems.

Lemma 6.14 *If F is an L_n-minimal 1-frame s.t. $|F| = L_n \cup X$ (where $L_n \cap X = \emptyset$) then*

$$\forall p, q \in |F|,\ F \not\models (p \rightarrow \mathrm{Neg}_p(q)) \wedge (q \rightarrow \mathrm{Neg}_q(p))$$

Proof. First we relabel the elements of X so that $|F| = L_{n+k}$, where $k = |X|$.

Consider $p, q \in |F|$. If $p \notin N_F(q)$ then trivially $F \not\models p \rightarrow \mathrm{Neg}_p(q)$, since $v(p) = v(q) = 1$ is consistent with F. The same also holds trivially if $N_F(q) = \{p\}$, since then $\mathrm{Neg}_p(q) = \bot$, or if $p = q$ since then $\mathrm{Neg}_p(p) = p$.

Now consider the case where $p \neq q, \{p\} \subset N_F(q)$. Construct the 1-frame

F' by deleting the edge $\{p, q\}$ from F. Since F is L_n-minimal we have $F' \not\lesssim_n F$. By Corollary 6.12 we then have $F' \not\lesssim_{n+k} F$. Hence at least one of the following is true

$$\exists w \in \mathrm{Val}\left(L_{n+k}\right), \ \left\|w\right\|_F > \left\|w\right\|_{F'} \tag{6.28}$$

$$\forall v \in \mathrm{Val}\left(L_{n+k}\right), \ \left\|v\right\|_F \geq \left\|v\right\|_{F'} \tag{6.29}$$

Suppose (6.29) holds. But there is a valuation $v \in \mathrm{Val}\left(F'\right)$ for which $v(p) = v(q) = 1$, and for any such valuation $v \notin \mathrm{Val}\left(F\right)$. Hence for such a v,

$$\left\|v\right\|_{F'} > 0 = \left\|v\right\|_F$$

which contradicts (6.29). Hence (6.28) must hold.

So choose a valuation w that satisfies (6.28). Then by Lemma 6.13, either

$$w(p) = 1 \text{ and } w(q) = w(\mathrm{Neg}_p(q)) = 0$$

or

$$w(q) = 1 \text{ and } w(p) = w(\mathrm{Neg}_q(p)) = 0$$

Whichever of the above holds the claim is proved, since in each case $w \in \mathrm{Val}\left(F\right)$ and

$$w\big((p \rightarrow \mathrm{Neg}_p(q)) \wedge (q \rightarrow \mathrm{Neg}_q(p))\big) = 0 \qquad \square$$

Lemma 6.15 *If F is an L_n-minimal 1-frame s.t. $|F| = L_n \cup X$ (where $L_n \cap X = \emptyset$) then for all $p \in X$ there is some $q \in N_F(p)$ s.t.*

$$F \not\models p \rightarrow \mathrm{Neg}_p(q)$$

Proof. Pick $p \in X$. Relabel X so that $|F| = L_{n+k}$ and $p = p_{n+k}$, where $k = |X|$. We can assume that for all $q \in N_F(p)$, $N_F(q) \setminus \{p\} \neq \emptyset$ since otherwise $\mathrm{Neg}_p(q) \equiv \bot$ and so the claim follows trivially. Now obtain F' from F by deleting p - F' is a valid 1-frame since it's corresponding graph has no isolated vertices.

Now, since F is L_n-minimal, $F' \not\preceq_n F$. Hence by Lemma 6.11, $F' \not\preceq_{n+k-1} F$. Therefore

$$\exists v \in \mathrm{Val}\,(L_{n+k-1}) \text{ s.t. } \|v\|_F > \|v\|_{F'}$$

But since $\|v\|_F \leq 1$, we see that $\|v\|_{F'} = 0$. Hence any such v must be of the form

$$v(p') = 0 \quad \forall p' \in N_F(p)$$

$$v(Neg_p(q)) = 0 \text{ for some } q \in N_F(p)$$

and for any $w \in \mathrm{Val}\,(L_{n+k})$ s.t. $w \upharpoonright L_{n+k-1} = v$ we have $w(p) = 1$. Hence for each $p \in X$ there is $q \in N_F(p)$ s.t.

$$F \not\models p \to \mathrm{Neg}_p(q) \qquad \square$$

We go on to extend this result to *all* vertices surrounding a vertex outside L_n.

Lemma 6.16 *Let F be an L_n-minimal 1-frame s.t. $|F| = L_n \cup X$ where $L_n \cap X = \emptyset$. Then for all $p \in X$, for all $q \in N_F(p)$*

$$F \not\models p \to \mathrm{Neg}_p(q)$$

Proof. Consider $p \in X$ and $q \in N_F(p)$. First notice that we can take, wolog, $N_F(q)\setminus\{p\} \neq \emptyset$, since in this case $\mathrm{Neg}_p(q) \equiv \bot$ and so the claim is trivial for such a p and q. Also, if $N_F(p) = \{q\}$ then the result holds by Lemma 6.15. Now, relabel X as $L_{n+k} \setminus L_n$ so that $p = p_{n+k}$, where $k = |X|$.

Suppose now that $N_F(p) \setminus \{q\} \neq \emptyset$. Let F' be obtained from F by deleting the edge $\{p, q\}$. Then since F' is L_n-minimal $F' \not\preceq_n F$, and by Corollary 6.12 we then have $F' \not\preceq_{n+k-1} F$. That is, there is some $v \in \mathrm{Val}\,(L_{n+k-1})$ s.t. $\|v\|_F = 1$ and $\|v\|_{F'} = 0$. Then for any such v, by Lemma 6.13, either

1. $v(q) = 1$ and $v(\mathrm{Neg}_q(p)) = 0$, or
2. $v(q) = 0$, $v(\mathrm{Neg}_q(p)) = 0$ and $v(Neg_p(q)) = 0$.

Now, if the first case holds then for the $w \in \mathrm{Val}\,(L_{n+k})$ s.t. $w \restriction L_{n+k-1} = v$ and $w(p) = 1$ we have $w \in \mathrm{Val}\,(F')$. Then $\|v\|_{F'} = 1$, which contradicts $\|v\|_{F'} = 0$. Hence the second case must hold.

So let $w \in \mathrm{Val}\,(L_F)$ be s.t. $w \restriction L_{n+k-1} = v$. Then we must have $w(p) = 1$, and so

$$F \not\models p \to \mathrm{Neg}_p(q)$$

as required. □

Lemma 6.17 *Let F be an L_n-minimal 1-frame s.t. $|F| = L_n \cup X$ where $L_n \cap X = \emptyset$. Then for all $p \in X$, for all $p' \notin N_F(p)$ there is some $q \in N_F(p)$ s.t.*

$$F \not\models (p \wedge p') \to \mathrm{Neg}_p(q)$$

Proof. 1. For $p' = p$ the claim is proved by Lemma 6.15.

2. Let $p \neq p'$. Relabel X as $L_{n+k} \setminus L_n$ where $k = |X|$ so that $p = p_{n+k}$. Form F' from F by adding the edge $\{p, p'\}$. Now since F is L_n-minimal we have $F' \not\preceq_n F$. Then by Corollary 6.12 we have $F' \not\preceq_{n+k-1} F$. That

is, there is some $v \in \mathrm{Val}\,(L_{n+k-1})$ s.t. $\|v\|_{F'} = 0$ and $\|v\|_F = 1$. Clearly any such v must have

(a) $v(p') = 1$

(b) $v(\mathrm{Neg}_{p'}(p)) = 0$

(c) There is some $q \in N_F(p)$ s.t. $v(\mathrm{Neg}_p(q)) = 0$

Notice that for any $w \in \mathrm{Val}\,(F)$ s.t. $w \restriction L_{n+k-1} = v$, we have $w(p) = 1$ and so $w \notin \mathrm{Val}\,(F')$. Hence

$$F \not\models (p \wedge p') \to \mathrm{Neg}_p(q)$$

as required. □

We now give a logical characterisation of the L_n-minimal 1-frames of size n.

Theorem 6.18 *If $|F| = L_n$ then F is L_n-minimal iff for all $p, q \in L_n$*

$$F \not\models (p \to \mathrm{Neg}_p(q)) \wedge (q \to \mathrm{Neg}_q(p))$$

Proof.

⇒ Immediate from Lemma 6.14.

⇐ Suppose for all p, q that

$$F \not\models (p \to \mathrm{Neg}_p(q)) \wedge (q \to \mathrm{Neg}_q(p))$$

Then for any p, q s.t. $p \in N_F(q)$ there is some $v \in \mathrm{Val}\,(F)$ s.t. either

1. $v(p) = 1$ and $v(q) = v(\mathrm{Neg}_p(q)) = 0$, or
2. $v(q) = 1$ and $v(p) = v(\mathrm{Neg}_q(p)) = 0$.

In either case removing the link $\{p, q\}$ gives a 1-frame F' for which v is not valid. Hence $F' \not\leq_n F$.

Also, if $p \notin N_F(q)$ then the valuation $v(p) = v(q) = 1$ is not consistent with adding the link $\{p, q\}$ and so any 1-frame F' so constructed has $F' \not\leq_n F$.

Hence any 1-frame F' formed from F by adding or removing an edge has $F' \not\leq_n F$, and hence F is L_n-minimal. □

Corollary 6.19 *For $|F| = L_n$, F is* ***not*** *L_n-minimal iff there is some $p, q \in F$ with $p \in N_F(q)$ and for which*

$$F \models (\mathrm{Neg}_p(q) \leftrightarrow q) \wedge (\mathrm{Neg}_q(p) \leftrightarrow p)$$

Remark: The previous Theorem and Corollary can be interpreted as saying that an L_n-minimal 1-frame on L_n has no "redundant" incompatibilities. Indeed, the Corollary can be read as saying that a 1-frame is *not* L_n-minimal iff there is some incompatibility p, q which may as well be removed - it will not affect the 1-frame. That is, p never forces q to be false unless some other variable is already forcing the same - and vice versa. In this sense, we may as well ignore the incompatibility between p and q.

This is a reassuring interpretation of L_n-minimality - it is a concept which selects as special those 1-frames which are most "efficient," in terms of their incompatibility structure.

We can also extend the previous result to 1-frames whose size is $n + 1$.

Theorem 6.20 *If $|F| = L_{n+1}$ then F is L_n-minimal iff all of the following hold:*

1. *For all* $p, q \in L_{n+1}$,

$$F \not\models (p \rightarrow \mathrm{Neg}_p(q)) \wedge (q \rightarrow \mathrm{Neg}_q(p)) \tag{6.30}$$

2. *For all* $p \in N_F(p_{n+1})$

$$F \not\models p_{n+1} \rightarrow \mathrm{Neg}_{p_{n+1}}(p) \tag{6.31}$$

3. *For all* $p \notin N_F(p_{n+1})$ *there is some* $q \in N_F(p_{n+1})$ *s.t.*

$$F \not\models (p \wedge p_{n+1}) \rightarrow \mathrm{Neg}_{p_{n+1}}(q) \tag{6.32}$$

Proof. $\Rightarrow$

1. By Lemma 6.14
2. By Lemma 6.16
3. By Lemma 6.17

$\Leftarrow$ Suppose F is s.t. (6.30), (6.31) and (6.32) above hold. Suppose F' is s.t. $|F'| = L_n$ and $F' \prec_n F$. As in the proof of Theorem 6.18 every edge of F' must be an edge of F. Consider now some $p \in N_F(p_{n+1})$. Then by (6.31) there is $v \in \mathrm{Val}\,(F)$ s.t. $v(p_{n+1}) = 1$, $v(p) = 0$ and $v(\mathrm{Neg}_{p_{n+1}}(p)) = 0$.

Let now $w = v \restriction L_n$. Then $\|w\|_F = 1$. But since $N_{F'}(p) \subset N_F(p)$ we have

$$w(\bigvee N_{F'}(p)) = v(\bigvee N_{F'}(p)) \leq v(\bigvee N_F(p)) = v(\mathrm{Neg}_{p_{n+1}}(p)) = 0$$

Hence $w \notin \mathrm{Val}\,(F')$ and so $\|w\|_{F'} = 0 < \|w\|_F$. Therefore $F' \not\prec_n F$.

So any 1-frame F' s.t. $F' \prec_n F$ must have $|F'| = L_{n+1}$.

Now, suppose $F' \prec_n F$ with $|F'| = L_{n+1}$. Again as in the proof of Theorem 6.18, F' has exactly the same edges $\{p, q\}$ where $p, q \in L_n$ as F.

Suppose now that F' has some edge $\{p, p_{n+1}\}$ which F does not. By (6.32) there is some $q \in N_F(p_{n+1})$ and a valuation $v \in \mathrm{Val}\,(F)$ s.t.

$$\begin{aligned} v(p) = 1 \quad & v(p_{n+1}) = 1 \\ v(q) = 0 \quad & v(\mathrm{Neg}_{p_{n+1}}(q)) = 0 \end{aligned}$$

Let $v' = v \restriction L_n$. Then $\|v'\|_F = 1$. Clearly $v \notin \mathrm{Val}\,(F')$ since $v(p) = v(p_{n+1}) = 1$ and $p \in N_{F'}(p_{n+1})$. Let $w \in \mathrm{Val}\,(L_{n+1})$ be s.t. $w \restriction L_n = v'$ and $w(p_{n+1}) = 0$. Then

$$\|v'\|_{F'} = \|v\|_{F'} + \|w\|_{F'} = \|w\|_{F'}$$

But since $w(q) = 0$ and $w(\bigvee N_{F'}(q)) = 0$, we have $w \notin \mathrm{Val}\,(F')$ and so $\|w\|_{F'} = 0$. Hence $\|v'\|_{F'} = 0 < 1 = \|v'\|_F$, and so $F' \not\prec_n F$.

So suppose that F has an edge $\{p, p_{n+1}\}$ which F' does not. By (6.31) there is some $v \in \mathrm{Val}\,(F)$ s.t. $v(p_{n+1}) = 1$, $v(p) = 0$ and $v(Neg_{p_{n+1}}(p)) = 0$. Let $v' = v \restriction L_n$, so that $\|v'\|_F = 1$.

But it is clear to see that $\|v'\|_{F'} = 0$. Indeed, for any $w \in \mathrm{Val}\,(L_{n+1})$ s.t. $w \restriction L_n = v'$ we have $w(p) = 0$ and

$$\begin{aligned} w(\bigvee N_{F'}(p)) &= w(\mathrm{Neg}_{p_{n+1}}(p)) \\ &= v(\mathrm{Neg}_{p_{n+1}}(p)) \\ &= 0 \end{aligned}$$

so that $w \notin \mathrm{Val}(F')$. Hence $F' \not\preceq_n F$. So there is no F' s.t. $F' \prec_n F$, and so F is indeed L_n-minimal. □

Remark: Again we can interpret the preceding theorem in a useful manner. Essentially it says that L_n-minimal frames on L_{n+1} are characterised by the following features:

1. They have no redundant incompatibilities, as in the Remark on Theorem 6.18.
2. The extra variable is not redundant - we cannot have an equally rich incompatibility structure without the extra variable p_{n+1}.
3. The third condition is more complicated to understand, but it can be read as claiming that the extra variable has "enough incompatibility." That is, adding further incompatibilities to the extra variable will result in redundancy in the 1-frames incompatibility structure.

Unfortunately, we again have no further results in this section.It seems that when we add 2 or more extra elements to L_n, logical characterisations become exceedingly hard to calculated.

6.8 The Behaviour of FE_1

To further explore the properties of the frame entropy inference processes we will study in detail the 1-dimensional case where the constraint set is defined on $L_1 = \{p_1\}$.

As stated in Section 6.4, we need only consider probability functions which are consistent with 1-frames whose non-trivial[2] part is L_1-minimal, and whose trivial

[2] For this section, we will consider the non-trivial part of a 1-frame to be the connective closure of L_1, and the trivial part to be the rest of the 1-frame

part is a Moon-Moser extremal graph. There are precisely three non-trivial 1-frames, namely K_2, K_3 and K_4, as shown in Section 6.5[3].

Now, since we wish to examine how the three different non-trivial parts will interact, we must consider FE_1^k for $k \geq 6$, since for $k < 5$ not all three are possible. We proceed by defining a maximum entropy function for each of K_2, K_3 and K_4. Let

$$\begin{aligned} T_1^k &= K_2 \cup M_{k-2} \\ T_2^k &= K_3 \cup M_{k-3} \\ T_3^k &= K_4 \cup M_{k-4} \end{aligned}$$

where in each case p_1 is an element of the complete graph K_i, and M_r is a Moon-Moser extremal graph of order r. Now set

$$\begin{aligned} f_1^k(x) &= \left(\tfrac{x}{m(k-2)}\right)^{-x} \left(\tfrac{1-x}{m(k-2)}\right)^{-(1-x)} \\ f_2^k(x) &= \left(\tfrac{x}{m(k-3)}\right)^{-x} \left(\tfrac{1-x}{2m(k-3)}\right)^{-(1-x)} \\ f_3^k(x) &= \left(\tfrac{x}{m(k-4)}\right)^{-x} \left(\tfrac{1-x}{3m(k-4)}\right)^{-(1-x)} \end{aligned}$$

Then $f_i^k(x)$ is the maximum entropy of a probability function w on L_k which is consistent with T_i and for which $w(p_1) = x$. Hence, $\mathrm{FE}_1^k(\Sigma)$ is that $w \in V^1(\Sigma)$ for which

$$\max\left\{f_1^k(x), f_2^k(x), f_3^k(x)\right\}$$

is maximal, where $x = w(p_1)$.

[3] Since there are no 1-frames of order 1, FE_1^1 is undefined and so we will only look at FE_1^k for $k \geq 2$.

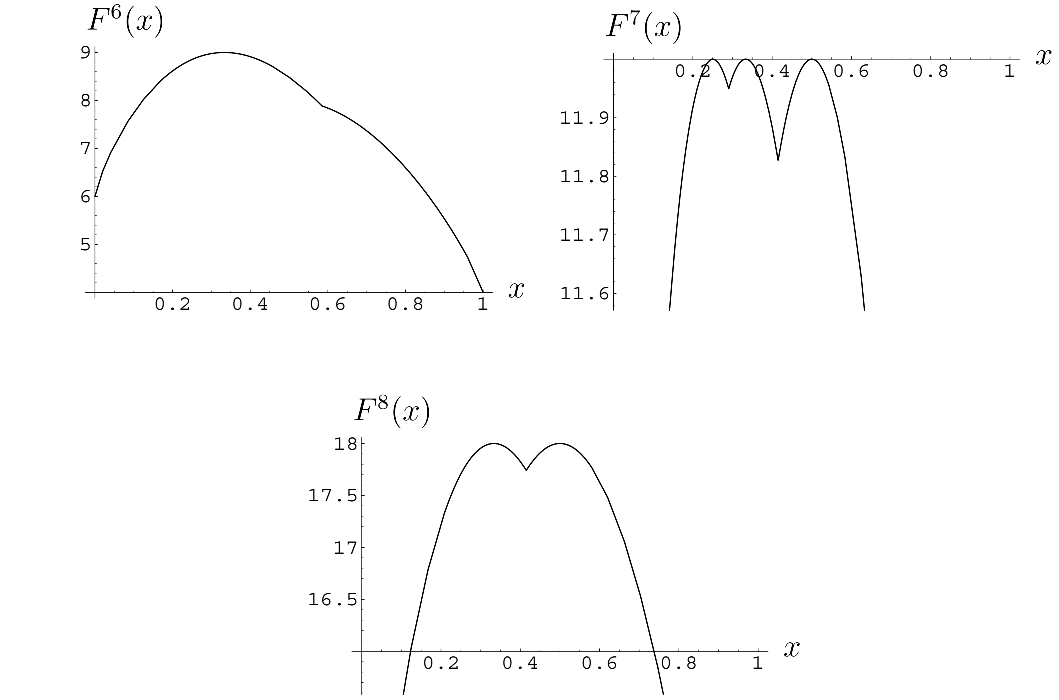


Figure 6.1: $F^k(x)$ for $k = 6, 7, 8$

So, now let $F^k(x) = \max\left\{f_1^k(x), f_2^k(x), f_3^k(x)\right\}$. It is clear that F^k depends on the Moon-Moser function $m\,(k)$, which behaviour varies awkwardly with k mod 3. However, we can easily see that for each f_i,

$$f_i^{k+3} = 3.f_i^k$$

since $m\,(k+3) = 3.m\,(k)$. Hence, any maximum of F^k will also be a maximum of F^{k+3}, and vice versa, and so we need only examine FE_1^k for three values of k (pairwise distinct $\mod 3$).

Although we might hope that the three cases turn out to be equivalent, the graphs in Figure 6.1 show that this is not the case. We will examine the three separate cases to see how they behave in the simple case where $\Sigma = \{w(p_1) \in [a, b]\ \}$

- $\mathbf{k \equiv 0}$

  $\mathrm{FE}_1^k(\Sigma)(p_1)$ is that $x \in [a, b]$ for which $|x - \frac{1}{3}|$ is minimal. This is reminiscent of classical ME, which (in the 1-dimensional case) minimises the distance to $\frac{1}{2}$.

- $\mathbf{k \equiv 1}$

  In this case F^k has two minima and three maxima and hence does not always have a unique maximum on $[a, b]$. The maxima occur at $\frac{1}{4}$,$\frac{1}{3}$ and $\frac{1}{2}$, while the minima occur at $t_1 = \frac{\ln 9 - \ln 8}{\ln 3 - \ln 2} \approx 0.290$ and $t_2 = \frac{\ln 4 - \ln 3}{\ln 2} \approx 0.415$.

  Hence $\mathrm{FE}_1^k(\Sigma)$ will have a unique solution for $k \equiv 1 \mod 3$ iff:

  1. $[a, b]$ contains at most one of the points $\frac{1}{4}$, $\frac{1}{3}$ and $\frac{1}{2}$, and;
  2. If $t_1 \in [a, b]$ and $\frac{1}{4} < a \leq b < \frac{1}{3}$ then $f_3^k(a) \neq f_2^k(b)$, and;
  3. If $t_2 \in [a, b]$ and $\frac{1}{3} < a \leq b < \frac{1}{2}$ then $f_2^k(a) \neq f_1^k(b)$.

- $\mathbf{k \equiv 2}$

  F^k has two maxima and one minimum in this case, so again $\mathrm{FE}_1^k(\Sigma)$ does not necessarily have a unique solution. The maxima are at $\frac{1}{3}$ and $\frac{1}{2}$ while the minimum is again at t_2.

  So $\mathrm{FE}_1^k(\Sigma)$ will have a unique solution for $k \equiv 2 \mod 3$ iff:

  1. $[a, b]$ contains at most one of the points $\frac{1}{3}$ and $\frac{1}{2}$, and;
  2. If $t_2 \in [a, b]$ and $\frac{1}{3} < a \leq b < \frac{1}{2}$ then $f_2^k(a) \neq f_1^k(b)$.

So we see that in this case we get three different answers, depending on the value of $k \mod 3$. Also, note that the existence of multiple maxima in the $k \equiv 1, 2$ cases indicate that this process will behave highly discontinuously. Changing the values of a and b (in the definition of Σ) by small amounts in the region of these maxima will sometimes add extra solutions, and will sometimes destroy existing solutions.

6.9 Discussion

The frame entropy inference process as defined here seems to behave very erratically, as evidenced by the example considered in Section 6.8. Also, as we saw in Section 6.6 and Section 6.7 it is extremely difficult to calculate the L_n-minimal frames. We cannot say even that there are a finite number of L_n-minimal 1-frames, except in the 1-dimensional case. These problems mean that there is very little we can learn from this inference process, at least at the current stage of development.

However, all is not lost. The partial success of the two attempts at characterisation of the L_n-minimal 1-frames suggests that it may in fact be possible to formulate some process for finding such frames. It is to be hoped that the work described here may give some clue as to how to proceed in this matter.

Also, the dependence of the inference process upon the values of $k \mod 3$ may not turn out to be such a problem. If we could discover that there are only a finite number of L_n-minimal frames then presumably the values given by this inference process would depend upon the values of $kmod3$. Then we could define the inference process so as to give the union of these 3 sets of answers. Unfortunately, our current paucity of results concerning this inference process means this can only be conjecture though.

Chapter 7

Conclusions

We present here a summary of the work in this thesis, and discuss further research possibilities.

7.1 Positive Frames

The notion of a positive frame seems to capture successfully the intuitions concerning perception and negation discussed in Chapter 2. The different types of positive frame such as normal frames, 1 -frames, c-frame and general positive frames allow for subtle distinctions between the different philosophical concepts concerned. Certain of the results shed light on the relationships between the different types of positive frames, and the correspondence with (hyper)graphs is of major importance in the study of these structures.

It would be interesting to investigate the theory of positive frames further, looking for additional relationships between the various flavours of frame. In particular, it may be of value to consider variations of the fundamental principles, especially with respect to the criticisms levelled at the Classification Principle in Chapter 2.

7.2 The CFE Inference Process

The CFE inference process is one which is very simple to specify, however it is quite complicated to understand how it behaves. The characterisations of the process as a model of expert knowledge given in Chapter 4 give, I believe, a solid justification to the process. However, due to the technical difficulty involved in working with this process, we have only managed to give a description of CFE on a certain small class of constraint sets (i.e., the adamant constraints). It would certainly be of importance to extend these results to ethereal constraint sets also, as discussed in Section 4.5.

Another problem with the complexity of the CFE process is the difficulty in producing examples of its behaviour. Chapter 5 considers only three properties of inference processes, two of which do not hold for CFE. The most important, Language Invariance does hold though, and the failure of CFE to satisfy Renaming and Continuity is not surprising given the nature of its definition. Further research into understanding the detailed behaviour of CFE with more examples would certainly be valuable.

7.3 The FE Inference Process

Our definition of FE in Chapter 6 seems to behave even more erratically than CFE. Perhaps this should not be surprising given that the inspiration for FE involves dropping the principle of Conjunctive Closure, a principle which goes a long way toward making CFE possible to analyse. However, the concept of specifying an inference process upon a (finite?) class of graphs is an attractive one — if the problems in discovering the L-minimal graphs could be overcome it may open the door to a easily calculable inference process. It is unfortunate that we only have partial characterisations of the class of L-minimal graphs. It is hoped however that these very different approaches might be extended to further results about this class of graphs.

The behaviour of FE in the one-dimensional case suggests that although this inference process might behave in a confusing manner, it might exhibit some interesting behaviour if we could easily analyse it. The discussion at the end of Chapter 6 proposes a way in which this might be accomplished.

In summary then, we have developed two unorthodox inference processes from a philosophical approach to perception and negation. The complexity of these processes indicates that much further work may be necessary, into both the philosophical fundamentals of the theory and its technical implementation.

Bibliography

[1] J.C. Beall. Is the observable world consistent? *Australasian Journal of Philosophy*, 78(1):113–118, March 2000.

[2] J.C. Beall and M. Colyvan. Looking for contradictions. *Australasian Journal of Philosophy*, 79(4):564–569, December 2001.

[3] C. Berge. *Hypergraphs: Combinatorics of Finite Sets.* North-Holland Mathematical Library, Amsterdam, 1989.

[4] A. Bundy. Incidence calculus: a mechanism for probabilistic reasoning. In *Proceedings of the Workshop on Uncertainty in Artificial Intelligence*, pages 177–184, Los Angeles, 1985. UCLA.

[5] J.P. Cleave. *A Study of Logics.* Clarendon Press, Oxford, 1991.

[6] N. Cocchiarella. On the logic of natural kinds. *Philosophy of Science*, 43:202–222, 1976.

[7] P. Courtney. *Models of Belief and Inference Processes.* PhD thesis, Manchester University, Manchester, 1992.

[8] R.T. Cox. *The Algebra of Probable Inference.* John Hopkins Press, 1961.

[9] A.P. Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society*, 30:105–47, 1968.

[10] H.B. Enderton. *A mathematical introduction to logic.* Harcourt/Academix Press, San Diego, 2nd edition, 2001.

[11] P. Erdös. On some extremal problems in graph theory. *Israel Journal of Mathematics*, 3:113–116, 1965.

[12] H. Gaifmann and M. Snir. Probabilities over rich languages, testing and randomness. *Journal of Symbolic Logic*, 47:495–548, 1982.

[13] P. Gärdenfors. Frameworks for properties: Possible worlds vs. conceptual spaces. In L. Haaparanta, M. Kusch, and I. Niiniluoto, editors, *Language, Knowledge and Intentionality*, volume 49 of *Acta Philosphica Fenna.* Helsinki, 1990.

[14] P. Gärdenfors. Induction, conceprtual spaces and AI. *Philosophy of Science*, 57:78–95, 1990.

[15] P. Gärdenfors. Meanings as conceptual structures. In M. Carrier and P. Machamer, editors, *Mindscapes: Philosophy, Science and the Mind*, pages 61–86. Pittsburgh University Press, 1997.

[16] H. Granger. Aristotle's natural kinds. *Philosophy*, 64:245–247, 1989.

[17] J.R. Griggs, C.M. Grinstead, and D.R. Guichard. The number of maximal independent sets in a connected graph. *Discrete Mathematics*, 68:211–220, 1988.

[18] G.M. Hardegree. An approach to the logic of natural kinds. *Pacific Philosphical Quarterly*, 63:122–132, 1982.

[19] M.J. Hill, J.B. Paris, and G.M. Wilmers. Some observations on induction in predicate probabilistic reasoning. Unpublished: Manuchester University.

[20] C. Howson. *Hume's Problem: Induction and the Justification of Belief*. Oxford University Press, Oxford, 2000.

[21] D. Hyde. From heaps and gaps to heaps of gluts. *Mind*, 106:641–60, 1997.

[22] E.T. Jaynes. Information theory and statistical mechanics. *The Physical Review*, 106(4):620–630, May 1957. Also available at `http://bayes.wustl.edu/etj/node1.html`.

[23] E.T. Jaynes. Bayesian methods: General background. In J. Justice, editor, *Maximum entropy and Bayesian methods in applied statistics: Proceedings of the Fourth Maximum Entropy Workshop*, pages 1–25, Alberta, Canada, 1984. University of Calgary, Cambridge University Press.

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

[25] G.L. Kline. N.A. Vasil'ev and the development of many-valued logics. In A.-T. Tymieniecka, editor, *Contributions to Logic and Methodology in honor of J.M. Bocheński*, chapter 17, pages 315–326. North-Holland Publishing Company, Amsterdam, 1965.

[26] G.J. Klir and T.A. Folger. *Fuzzy Stes, Uncertainty and Information.* Prentice-Hall, New Jersey, 1988.

[27] S.A. Kripke. Identity and necessity. In M.K. Munitz, editor, *Identity and Individuation*, pages 135–164. New York University Press, New York, 1971.

[28] S.A. Kripke. *Naming and Necessity.* Harvard University Press, 1980.

[29] A.A. Mauer. Ockham's razor and chatton's anti-razor. *Mediaeval Studies*, 46:463–75, 1984.

[30] I. Maung. PhD thesis, Manchester University, Manchester, 1992.

[31] P. May. *Forecasting methods for horseracing.* Raceform Ltd., Newbury, England, 1998.

[32] J.W. Moon and L. Moser. On cliques in graphs. *Israel Journal of Mathematics*, 3:23–28, 1965.

[33] J. Paris. Common sense and maximum entropy. *Synthese*, 117:75–93, 1999.

[34] J.B Paris. *The Uncertain Reasoner's Companion: A Mathematical Perspective.* Cambridge University Press, Cambridge, England, 1994.

[35] J.B. Paris and A. Vencovská. On the applicability of maximum entropy to inexact reasoning. *International Journal of Approximate Reasoning*, 3:1–34, 1989.

[36] J.B. Paris and A. Vencovská. A note of the inevitability of maximum entropy. *International Journal of Approximate Reasoning*, 4:183–223, 1990.

[37] J.B. Paris, A. Vencovská, and G.M. Wilmers. A natural prior probability distribution derived from the propositional calculus. *Annals of Pure and Applied Logic*, 70:243–285, 1994.

[38] J.B. Paris, P.N. Watton, and G.M. Wilmers. On the structure of probability functions in the natural world. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 8(3):311–329, 2000.

[39] J. Pearl. On evidential reasoning in a hierarchy of responses. *AI*, 28:9–15, 1986.

[40] D.R. Pears. Incompatibilities of colours. In A.G.N. Flew, editor, *Logic and Language*, Second Series, pages 112–24. Blackwell, Oxford, 1953.

[41] G. Priest. Negation as cancellation, and connexive logic. *Topoi*, 18:141–148, 1999.

[42] G. Priest. Perceiving contradictions. *Australasian Journal of Philosophy*, 77(4):439–446, December 1999.

[43] G. Priest. Truth and contradiction. *Philosophical Quarterly*, 50(200):305–319, 2000.

[44] H. Putnam. The meaning of ”meaning”. In *Mind, Language and Reality*, pages 215–71. Cambridge University Press, Cambridge, 1975.

[45] W.V. Quine. Three grades of modal involvement. In *Proceedings of the XIth International Congress of Philosophy*, volume 14, pages 156–174, Brussells, 1953.

[46] W.V. Quine. Natural kinds. In *Ontological Relativity and other essays*. Columbia University Press, New York, 1969.

[47] J.-P. Sarte. *L'Etre et le Néant*. Gallimard, Paris, 1943.

[48] W. Shannon, C.E. an Weaver. *The mathematical theory of communication*. University of Illinois Press, Urbana, Illinois, 1964.

[49] E. Sosa and J. Kim, editors. *Epistemology: An Anthology*. Blackwell Philosphy Anthologies. Blackwell Publishers, 1999.

[50] D. Tversky, A. an d Kahneman. Judgement under uncertainty: Heuristics and biases. *Science*, (185):1124–1131, 1974.

[51] N.A. Vasil'ev. Logica i métalogica. *Logos*, 1–2:53–81, 1913. Translated into English by V. Vasukov as [52].

[52] V. Vasukov. Logic and metalogic. *Axiomathes*, 4, 1993. Translation into English of [51].

[53] T.E. Wilkerson. *Natural Kinds*. Aldershot, Avebury, 1995.

[54] C. Witteman and S. Renooij. Evaluation of a verbal-numerical probability scale. *International Journal of Approximate Reasoning*, 33:117–131, 2003.

[55] L. Wittgenstein. *Tractatus logico-philosophicus*. Routledge and Kegan-Paul, 1922.

Appendix A

Useful Results

Some useful arithmetical results and results from other papers are stated here to save space in the main thesis.

A.1 Arithmetical Results

Lemma A.1 *Let $x_1, x_2, \ldots, x_n$ be real numbers and let $y_1, y_2, \ldots, y_n$ be positive real numbers. Then*

$$\frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n y_i} \leq \sum_{i=1}^n \frac{x_i}{y_i}$$

Proof.

$$\frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n y_i} = \sum_{j=1}^n \frac{x_j}{\sum_{i=1}^n y_i} \leq \sum_{j=1}^n \frac{x_j}{y_j}$$

□

A.2 Maximum Entropy Results of Paris and Vencovská

The two theorems given here are Theorems 1 and 2 of [35], stated here for convenience. We first give a brief summary of the notation used in that paper.

As in subsection 4.3.2 of this thesis they assume the existence of a constraint set which is defined similarly to the expert constraint sets of Definition 4.8. The conditions (1) and (2) referred to are the equivalents of (4.8) and (4.9) — the conditions on the $\approx$ relationship. There is an example set M of size N with subsets $A_0, A_1, \ldots, A_{n-1}$. A "model of the constraints" is an assignment of $A_0, A_1, \ldots, A_{n-1}$ to specific subsets of M which satisfies the constraint set. $B(\mathbf{A})$ is a boolean combination of the subsets $A_1, A_2, \ldots, A_n$ and $g_1, g_2, \ldots, g_w$ number the "atoms" of $B(\mathbf{A})$. That is,

$$B(\mathbf{A}) = C_{g_1} \cup C_{g_2} \cup \cdots \cup C_{g_w}$$

where $C_1, C_2, \ldots, C_{2^n}$ are the subsets of M formed by taking the intersections

$$C_f = \bigcap_{i<n} A_i^{f(i)}$$

over all functions $f : \{0, 1, \ldots, n-1\} \to \{0, 1\}$ where $A_i^1 = A_i$ and $A_i^0 = M \setminus A_i$. ρ is the maximum entropy solution of the limit form of the constraint set, and ρ_i denotes the value of ρ in its ith coordinate. The theorems are:

Theorem A.2 *Assume the given constraints are consistent. Then for each $\mu, \nu > 0$ there exist N_0 and $\epsilon > 0$ such that for all $N \geq N_0$ and $\approx$ satisfying (1)* *and (2) the proportion of the models of the constraints for which*

$$\left| \sum_{i=1}^{w} \rho_{g_i} - \frac{B(\mathbf{A})}{N} \right| \geq \nu$$

is at most μ.

That is, if M is taken sufficiently large the majority of the models of the constraints will define a probability function arbitrarily close to the maximum entropy solution.

Theorem A.3 *If [the limit form of the constraints] and $\sum_{i=1}^{w} p_{g_i} > 0$ has a solution, then the maximum entropy solution ρ of [the limit form of the constraints] also satisfies $\sum_{i=1}^{w} \rho_{g_i} > 0$.*

This theorem states that the maximum entropy solution will not give zero probability to any proposition unless forced to by the constraints.

A.3 Properties of the Moon & Moser and Griggs functions

Lemma A.4 *Let r and s be natural numbers. Then*

$$m\,(r)\,.m\,(s) \leq m\,(r+s) \ \text{ for } r, s \geq 0, \quad \text{ and}$$
$$m\,(3r)\,.m\,(s) = m\,(3r+s) \ \text{ for } r \geq 0, s \geq 2.$$

Proof. • Firstly, let G_1 and G_2 be two graphs of order r and s respectively, which have $m\,(r)$ and $m\,(s)$ maximal independent sets. Then $G_1 \cup G_2$ is a graph of order $r+s$ which has $m\,(r)\,.m\,(s)$ maximal independent sets, since any maximal independent set X of $G_1 \cup G_2$ is of the form $X = X_1 \cup X_2$ where X_i is a maximal independent set of G_i.

However, the maximum number of maximal independent sets that a graph of order $r + s$ may have is $m\left(r+s\right)$, and so $m\left(r\right).m\left(s\right) \leq m\left(r+s\right)$.

- The second part of the Lemma is by induction and is obvious from inspection of m — for $s \geq 2$, it is obvious that

  $$m\left(3\right).m\left(s\right) = 3.m\left(s\right) = m\left(s+3\right)$$

  Hence, by induction, for $r \geq 1$

  $$\begin{aligned} m\left(3r+s\right) =& 3.m\left(3(r-1)+s\right) \\ =& 3.m\left(3(r-1)\right).m\left(s\right) \\ =& m\left(3r\right).m\left(s\right) \end{aligned}$$

□

Corollary A.5 *For $r \geq s \geq 1$,*

$$m\left(r-s\right) \leq \frac{m\left(r\right)}{m\left(s\right)}$$

Proof. Immediate from Lemma A.4 with r replaced with $r - s$. □

Lemma A.6 *For $n \geq 4$,*

$$m\left(n-2\right) + m\left(n-3\right) + 2 > g\left(n\right)$$

$$2.m\left(n-3\right) + m\left(n-4\right) + 2 > g\left(n\right)$$

Proof. 1. $\mathbf{m}\left(\mathbf{n}-\mathbf{2}\right)+\mathbf{m}\left(\mathbf{n}-\mathbf{3}\right)+\mathbf{2}>\mathbf{g}\left(\mathbf{n}\right)$

The claim is easily checked by inspection for $4 \leq n \leq 11$. Now, for $n \geq 12$ let

$$x(n)=\frac{m\left(n-2\right)+m\left(n-3\right)}{g\left(n\right)}$$

We examine $x(n)$ for $n \equiv 0,1,2 \mod 3$

- $\mathbf{n \equiv 0}$

$$\begin{aligned} x(n) &= \frac{4.3^{\frac{n-6}{3}}+3^{\frac{n-3}{3}}}{2.3^{\frac{n-3}{3}}+2^{\frac{n-3}{3}}} \\ &= \frac{2.3^{\frac{n-3}{3}}+3^{\frac{n-6}{3}}}{2.3^{\frac{n-3}{3}}+2^{\frac{n-3}{3}}} > 1 \end{aligned}$$

- $\mathbf{n \equiv 1}$

$$\begin{aligned} x(n) &= \frac{2.3^{\frac{n-4}{3}}+4.3^{\frac{n-7}{3}}}{3^{\frac{n-1}{3}}+2^{\frac{n-4}{3}}} \\ &= \frac{3^{\frac{n-1}{3}}+3^{\frac{n-7}{3}}}{3^{\frac{n-1}{3}}+2^{\frac{n-4}{3}}} > 1 \end{aligned}$$

- $\mathbf{n \equiv 2}$

$$\begin{aligned} x(n) &= \frac{3^{\frac{n-2}{3}}+2.3^{\frac{n-5}{3}}}{4.3^{\frac{n-5}{3}}+3.2^{\frac{n-8}{3}}} \\ &= \frac{4.3^{\frac{n-5}{3}}+3^{\frac{n-5}{3}}}{4.3^{\frac{n-5}{3}}+3.2^{\frac{n-8}{3}}} > 1 \end{aligned}$$

Hence $x(n) > 1$ for all $n \geq 12$, and so the claim is proved.

2. $\mathbf{2.m}\left(\mathbf{n}-\mathbf{3}\right)+\mathbf{m}\left(\mathbf{n}-\mathbf{4}\right)+\mathbf{2}>\mathbf{g}\left(\mathbf{n}\right)$

Again, the claim can be easily checked by inspection for $4 \leq n \leq 7$. For $n \geq 8$, set

$$y(n)=\frac{2.m\left(n-3\right)+m\left(n-4\right)}{g\left(n\right)}$$

We examine $y(n)$ for $n \equiv 0, 1, 2 \mod 3$

- $\mathbf{n \equiv 0}$

$$y(n) = \frac{2.3^{\frac{n-3}{3}} + 2.3^{\frac{n-6}{3}}}{2.3^{\frac{n-3}{3}} + 2^{\frac{n-3}{3}}} > 1$$

- $\mathbf{n \equiv 1}$

$$\begin{aligned} y(n) &= \frac{8.3^{\frac{n-7}{3}} + 3^{\frac{n-4}{3}}}{3^{\frac{n-1}{3}} + 2^{\frac{n-4}{3}}} \\ &= \frac{3^{\frac{n-1}{3}} + 2.3^{\frac{n-7}{3}}}{3^{\frac{n-1}{3}} + 2^{\frac{n-4}{3}}} > 1 \end{aligned}$$

- $\mathbf{n \equiv 2}$

$$y(n) = \frac{4.3^{\frac{n-5}{3}} + 4.3^{\frac{n-8}{3}}}{4.3^{\frac{n-5}{3}} + 3.2^{\frac{n-8}{3}}} > 1$$

Hence $y(n) > 1$ for all $n \geq 8$, and so the claim is proved.

$\square$