View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Helsingin yliopiston digitaalinen arkisto

Ruurik Holm

Constructive Perspectives on Inductive Logic

Academic dissertation to be publicly discussed, by due permission of the Faculty of Arts at the University of Helsinki in auditorium XV, on the 20th of November, 2009 at 12 o'clock.

ISBN 978-952-92-6352-3 (paperback)
ISBN 978-952-10-5835-6 (PDF)
Helsinki 2009
Helsinki University Print

Abstract

Constructive (intuitionist, anti-realist) semantics has thus far been lacking an adequate concept of truth in infinity concerning factual (i.e., empirical, non-mathematical) sentences.

One consequence of this problem is the difficulty of incorporating inductive reasoning in constructive semantics. It is not possible to formulate a notion for probable truth in infinity if there is no adequate notion of what truth in infinity is. One needs a notion of a constructive possible world based on sensory experience. Moreover, a constructive probability measure must be defined over these constructively possible empirical worlds.

This study defines a particular kind of approach to the concept of truth in infinity for Rudolf Carnap's inductive logic. The new approach is based on truth in the consecutive finite domains of individuals. This concept will be given a constructive interpretation. What can be verifiably said about an empirical statement with respect to this concept of truth, will be explained, for which purpose a constructive notion of epistemic probability will be introduced.

The aim of this study is also to improve Carnap's inductive logic. The study addresses the problem of justifying the use of an "inductivist" method in Carnap's λ-continuum. A correction rule for adjusting the inductive method itself in the course of obtaining evidence will be introduced. Together with the constructive interpretation of probability, the correction rule yields positive prior probabilities for universal generalizations in infinite domains.

Contents

Acknowledgements

Most of this study has been conducted at the Department of Philosophy of the University of Helsinki, where my supervisor Ilkka Niiniluoto has assisted me greatly all the way through the process. Without his expertise, support and confidence in my work, it would never have been accomplished.

The Faculties of Philosophy at Utrecht University and at the University of Groningen provided me an international research community for studying the fields of constructive semantics and inductive logic. Albert Visser in Utrecht introduced me to the academic life in the Netherlands. Mark van Atten and Igor Douven were certainly among the most inspiring roommates a PhD student in this subject can have. During my stay in Groningen, Theo Kuipers provided me valuable guidance on various aspects of my work.

Already before I started my PhD research, I had become interested in constructive logic, especially in its type-theoretical variant. The research group on type theory at the Department of Philosophy in Helsinki was my first touch with top class academic research. The influence of my first supervisor Aarne Ranta on my philosophical thinking has been significant. Jan von Plato and Petri Mäenpää of the same research group have also supported me to a great extent in terms of both physical research facilities and intellectual training.

Gabriel Sandu has been, all along my postgraduate studies, more a colleague and a friend than "merely" a professor. His willingness to help and advise less experienced scholars in their academic efforts is remarkably generous.

The examiners of my thesis manuscript, Theo Kuipers, Per-Erik Malmnäs and Juha Oikkonen provided such detailed remarks and suggestions on the text that their importance to the final form of the thesis cannot be exaggerated.

In addition to his comments on some early drafts of my text, discussions with Aki Lehtinen on various mathematical topics in the kitchen of the Department of Philosophy have been most fascinating and inspiring. Juha Himanka, Timo Kaitaro, Petter Korkman and Markku Roinila, "the lunch bunch", has been a regular source of philosophical recreation.

As regards the administrative issues, Auli Kaipainen and Terhi Kiiskinen have greatly facilitated my presence at the Department of Philosophy in Helsinki.

I wish to extend my thanks also to Dag Prawitz, Jan Bergstra, Janne Hiipakka, Juha Ruokolainen, Matti Kinnunen, Roope Lenkkeri, Mirja Hartimo, Jan-Willem Romeyn, Matti Sintonen, Panu Raatikainen, Ondrej Majer, Jouko Väänänen and Juliette Kennedy.

The work on this study has been funded by the Leo and Regina Wainstein Foundation, the Academy of Finland, the Emil Aaltonen Foundation, the research group on type theory at the University of Helsinki, the research project *Logic – Theory and Applications* at the Departments of Philosophy and Mathematics of the University of Helsinki, the Oskar Öflund Foundation, the Otto A. Malm Foundation, the Department of Philosophy at the University of Helsinki and the University of Helsinki grant for finalizing a dissertation.

The last phase of the work took place while I have for many years been employed outside the research community. I wish to express my gratitude to my current employer in granting the required leaves of absence.

In the spirit of the antiauthoritarian 60's, my parents have been wise not to suggest a direction in my life; instead, they have supported me on the chosen path.

Niina's patience and understanding during the last hectic stage of the work has been indispensable.

This thesis is dedicated to my daughter Saaga.

Helsinki, October 2009

Ruurik Holm

Chapter 1

Introduction

1.1 General rationale and background

Constructive (intuitionist, antirealist) semantics has thus far lacked an adequate concept of truth in infinity concerning factual (i.e., empirical, non-mathematical) sentences. The problem has been how to define a verification condition for sentences that can neither be mathematically proved nor verified by observations, like universal generalizations in infinite non-mathematical domains of individuals?

One consequence of this problem is the difficulty to incorporate inductive reasoning in constructive semantics. For probabilistic induction concerning statements in an infinite domain of individuals, one needs a concept of probability in infinity – and the probability of a statement in infinity is usually explained as the probability of its truth in infinity.

Benenson (1984), for example, provides arguments showing why constructive probability should necessarily be epistemic probability. Without dwelling on this discussion, the present study will simply be limited to epistemic probability.

Carnap's inductive logic, whose main reference is Carnap (1962), has lacked a sufficient explication of truth in infinity. In inductive logic, probability in infinity is defined as the limit of probabilities in consecutive finite cardinalities, which means that the concept of truth involved in the definition of probability is truth in a finite domain of individuals. Hence, a precise connection between probability in consecutive infinite domains and any concept of truth in infinity is missing from the system.

This study defines a particular kind of approach to the concept of truth in infinity for Carnap's inductive logic. This new approach is based on truth in consecutive finite domains of individuals. This concept will be given a constructive interpretation, and what can be verifiably said about an empirical statement with respect to this concept of truth, will be explained. For this purpose a constructive

notion of epistemic probability will be introduced.

Now the reader is entitled to ask why Carnap's inductive logic should be considered as the basic framework for defining constructive probabilities. Why not some other approach to defining epistemic probability? It is not enough to state that the study will be about a constructive variant of Carnap's inductive logic. One must offer some justification for building such a variant.

An important justification is the restriction on truth and probability of a formal language. Carnap's state description semantics is the best-known example of trying to define the concept of probability by using logical semantics. This approach is intuitively appealing, since the models of the semantics (state descriptions in Carnap's system) define the space of possibilities. Probability is most naturally defined by assigning probabilities to these entities representing the possible worlds.

Constructively possible worlds can be given as sequences of observations or data streams. One can represent such a data stream with binary sequences which possibly have no upper limit, each digit corresponding to an atomic fact. This is precisely what Carnap's state descriptions must be constructively: sequences of atomic facts, obtained one after another. Hence, the philosophy behind this definition of a constructive possible world is closely connected to Carnap's state descriptions. The constructive definitions can be built on the already existing system of state description semantics.

Still another advantage of Carnap's inductive logic is that its concept of probability is defined by using consecutive finite logical spaces without infinite sequences or logical spaces. This finitistic concept of probability is a suitable basis for defining probabilities for observation sequences. Carnap's inductive logic is thus in this sense based on constructively solid foundations. Moreover, there is a connection between the problem of induction and constructive semantics which makes formalization with finite logical spaces even more appropriate (see section 1.2 below).

Using Carnap's inductive logic as a background framework also means limiting the scope of the study. Carnap's systems in 1962 and 1952 are defined for monadic languages only. Many of the findings apply only to the variant of Carnap's inductive logic developed in this study. On the other hand, this is not necessarily a drawback, since the objective has been to provide merely an explication of constructive factual truth and constructive epistemic probability, not to address the whole field of research concerning these notions. Moreover, the study brings insight into the general problems behind defining an epistemic concept of probability and constructive probability. One can even establish a certain connection between these two concepts.

This study does not work within any particular mathematical system of constructive semantics. One system that may be conceivable for such a purpose is Per

Martin-Löf's constructive type theory and its nonstandard extension. However, it has become interesting enough to discuss the general philosophy of empirical constructive truth and probability and to sketch how the concepts could be worked into Carnap's state description semantics. The philosophical background to explaining these concepts had to be discussed in any case to justify the use of a particular mathematical system. It is a matter for a further study to actually construct a complete semantics on the basis of the preliminary discussions of this study.

The concept of constructivism which is used in this study is thus not defined in any formal system or using the meaning explanations pertaining to such a system. A precise concept of constructive empirical truth and probability would be impossible to start with since the very objective of the study is to define such concepts. However, the general principle of constructive truth is respected: there is no truth beyond knowledge (verification) and a sentence can only be true if it is knowable (verifiable). Constructive functions are assumed to be effectively computable – although a reference to a more relaxed concept of the Markov-constructive limit is provided in chapter 5. Finally, the standard BHK interpretation introduced by Heyting (1934) for logical constants will be used unless stated otherwise.

The previous literature on the probability of constructive empirical truth is slight. Benenson (1984) discusses an anti-realist (constructive) interpretation of probability statements themselves (i.e., not sentences to which probabilities are assigned) and holds that "logical relation theories" by Carnap (1962) and Keynes (1921) (i.e., inductive logic) provide a foundation for an anti-realist explanation of the meaning of probability statements. Benenson maintains that a realist account of probabilities is tantamount to an empirical account (for example, frequentist or propensity) (cf. Benenson 1984, p. 59), of which we have only probable knowledge, and thus we are in a regress in trying to explain the meaning of probability.

Non-empirical probabilities are, however, not necessarily anti-realist since they may exist without being known. As Grove, Halpern & Koller (1996, pp. 252-253, 264-273) have pointed out extending the results of Liogon'kii (1969), inductive probabilities are in general undecidable, but it is true that the logical relation theory *enables* the definition of an anti-realist concept of probability. At the same time it is possible to defend the view that the logical relations theory is objective as contrasted to the subjective interpretation of probability, which might also be eligible for an anti-realist interpretation.

With respect to constructive empirical truth, Benenson (1984, pp. 57-58) relies on Dummett's remark about the assertibility conditions for empirical statements. According to Dummett, for empirical statements "there will, for the anti-realist, be no question of there being anything in virtue of which they are (definitively) true, but only of things in virtue of which they are probably true". Dummett continues: "[...] and there is nothing to prevent a statement being so used that

we do not treat anything as conclusively verifying it." Dummett (1978, p. 162.) Dummett thus holds that no concept of conclusive verification is even needed for empirical statements. The key for Dummett is justified assertibility; a sentence may be justifiably asserted if evidence supporting it is obtained. This evidence does not have to verify the sentence conclusively, which entails that a justifiably assertible sentence may in the future not be justifiably assertible. On this view, the meaning of a sentence is defined by the condition of its justifiable assertibility, not verifiability.

Justified assertibility is connected with the notion of probability. Evidence increases the probability of a statement. A statement having at least a certain probability can be considered as justifiably assertible.

Evidence can thus make a sentence more probable and thus justifiably assertible, but what probability means must still be explained. Clearly the probability of a sentence means probability that the sentence holds, i.e., is true. If the domain of discourse – the domain of individuals – is infinite, then truth here means truth in infinity. Hence, it seems difficult to do away with the concept of constructive truth in infinity by replacing verifiability with justified assertibility. What "probably true" means in constructive terms needs to be defined, but Benenson (1984) does not provide an account of this.

1.2 Inductive inference

The aim of this study is also to improve Carnap's inductive logic, solving some of its traditional problems. The use of Carnap's inductive logic as the framework for the discussion about constructive probability needs some justification of the system of inductive logic itself – inductive logic must be a feasible explication of probability. To contest the criticism it has faced during the past decades, some arguments in favour of inductive logic will be provided in this study.

Scientific hypotheses are conventionally formalized as universally quantified sentences, universal generalizations. Because confirmation by evidence is a crucial question when dealing with scientific hypotheses, confirmation of universal generalizations is an important issue in the philosophy of science.

The problem arising from this can be formulated as follows.

Since there can be only a finite number of observations in a finite time, it is not possible to verify a universal generalization by means of evidence in an infinite domain of individuals. Hence, at best one can assign a positive probability to the universal generalization. The evidence e can confirm universal generalization h in

the sense of increasing its probability when the rule of conditional probabilities

$$P(h|e) = \frac{P(h\&e)}{P(e)} \tag{1.1}$$

is applied.

However, one should define a probability function by which this confirmation actually takes place, which has turned out to be difficult. Carnap's (1962) semantics of logical probability is an attempt to formalize the concept of probability by using the formal language of predicate logic, but it is well-known that in Carnap's system as well as in its generalization in Carnap (1952), universally quantified sentences are assigned a zero prior probability, which means that they must, according to the conditionalization rule above, be assigned zero probability under any finite body of evidence.

Scientific inference belongs to a more general framework of non-deductive or inductive inference, in which the conclusion does not follow logically from the premises. This type of inference is troubled by the problem of induction, which challenges any inference from past observations to future ones. For example, if one has observed n black crows, the inductive skepticism denies that one would be justified in concluding that the $n+1$'th crow is also black. Expressed in terms of probabilistic induction, inductive skepticism says that the probability of $n+1$'th crow's being black is not increased by evidence consisting of n black crows.

Consider now the inductive inference from the premiss that one has observed n black crows to the conclusion that $n+1$'th crow is also black. Moreover, suppose that this kind of simple enumerative induction is a valid form of inference. It follows that all the $n+1$ crows are black, from which one can conclude that $n+2$'th crow is also black, and so on.

In fact, it follows that all crows are black. Consider the meaning explanation of a universal quantification: $(\forall x)P(x)$ is true iff $P(c)$ is true for an arbitrary c. Hence, a valid inductive inference from a finite number of observations to the next instance entails the validity of the inference from these observations to a general law.

A conclusion about a single instance is at least possible to verify (or falsify) by observation when time passes. In a semantics based on verification conditions, it is possible to say that a single-instance conclusion of an inductive inference is indirectly verified by induction, meaning that the inference in question concludes that the instance will be directly observable. Moreover, if an arbitrary instance is directly observable, the corresponding universal law is verified and thus constructively true. Hence, if simple enumerative induction was a sound method of inference, empirical universal generalizations would have a method of verification in infinite domains of individuals; in other words, empirical universal generalizations would have a constructive meaning explanation. The rejection of the validity

of enumerative inductive inference thus has a central role in the problem of constructive semantics for empirical statements.

However, probabilistic induction attributing the probability p to the next instance does not provide a verification method for universal generalizations in the same fashion as above. The assumption that probabilistic induction is a valid form of inference does not solve the problem of constructive empirical truth in infinite domains.

The cardinality of the domain of individuals also links constructive semantics to the problem of induction. In the context of inductive inference, since a finite number of positive instances does not in general justify a universal law, a finite number of observed individuals does certainly not justify the conclusion that there is an infinite number of individuals in the domain. In fact, this conclusion can never be verified by observation. Hence, a way to formalize the domain of individuals which leaves open the possibility for the domain to be either finite or infinite would be appropriate both from the point of view of constructive semantics and inductive inference. This is one reason for implementing the semantics of this study on the basis of finite structures.

As was stated above, making an inductive inference concerning a thus far unobserved single individual is a different issue from assigning universal quantifications a positive prior probability. Even if observed black crows increase the probability that the next crow is black, universal generalizations may still have a zero prior probability.

On the other hand, if observed black crows do not increase the probability of the next crow's being black, the probability of an infinite series of black crows is obviously zero. Hence, at least some inductive non-skepticism is a prerequisite for positive probabilities of universal generalizations.

The way to tackle the zero prior problem of universal generalizations is obtained through an attempt to solve the problem of induction (see below).

This study addresses the problem of justifying the use of an "inductivist" method in Carnap's λ-continuum, which corresponds to the problem of induction in Carnap's framework. There is a particular method in the λ-continuum which does not allow of probabilistic induction. So far there has been no way of excluding such a non-inductivist method without inductive presumptions, i.e., without assuming in one way or another that induction is a valid method of inference.

The suggestion of introducing second-order probabilities (i.e., probabilities of inductive methods) for choosing among the inductive methods is rejected. Instead, a correction rule of adjusting the inductive method itself in the course of obtaining evidence will be introduced.

Together with the constructive interpretation of probability the correction rule yields positive prior probabilities for universal generalizations in infinite domains. This provides a solution to the classical zero prior problem referred to above.

Moreover, since it will be established that the probability of a sentence in an infinite domain actually corresponds to a particular notion of truth in infinity, universal generalizations can receive non-zero asymptotic probabilities while the probability in question is related to a notion of truth.

1.3 Outline of the study

The study is divided into the following chapters.

Chapter 2 of the study discusses what kinds of entities constructive empirical worlds are and how they can be linguistically denoted. The latter question also pertains to the problem of representing all the logically possible state descriptions.

Chapter 3 discusses the constructive formulation of the concept of epistemic probability. Special attention is devoted to the problem of formalizing probability and truth in the case of infinitary state descriptions.

Chapter 5 elaborates further the problems of constructive factual truth in infinity and probability. An attempt to resolve the problems by introducing the concepts of extendible truth and extendible probability is presented. Various properties of extendible truth and probability are displayed and proved. Section 5.3 brings up another hitherto overlooked problem related to asymptotic probabilities, which can also be solved by extendible truth.

Chapter 6 brings up the key problem of the logical interpretation of probabilities, namely their dependence on the choice of the prior distribution (i.e., inductive method in Carnap's terminology). Attempts to deal with the issue by introducing probabilities at the meta-level are scrutinized. The main outcome of the chapter is that second order probabilities do not provide any real help in justifying probabilistic inductive reasoning.

The arbitrariness in choosing the inductive method is met by a non-Bayesian way of adjusting the inductive method according to incoming evidence (ch. 7).

Section 7.2 first discusses a criterion based on the concept of immodesty for choosing among the inductive methods based on the performance of the method as evaluated by itself. It seems that this immodesty criterion is too liberal. In section 7.3, Carnap's own measure of success of an inductive method, the mean square error, is elaborated. Section 7.4 presents a non-Bayesian rule for updating the inductive methods, the correction rule Θ. The performance of the correction rule is evaluated in section 7.5. It is argued that a more general form of the correction rule performs in a particular sense better than a given non-optimum method of the λ-continuum.

The chapter concludes with a proof that a particular uniform and endless stream of evidence sentences has a positive limit probability when the correction rule is applied. This is one step toward establishing a positive probability for uni-

versal generalizations, which has been one of the problematic issues in inductive logic.

Finally, chapter 8 discusses the consequences which the adoption of a correction rule has in constructive semantics. It is argued that, *inter alia,* the prior probability of a particular universal generalization must be the same as the probability of the corresponding stream of evidence statements, in other words positive.

Chapter 2

Infinite possible worlds

This chapter discusses the concept of a possible world and the totality of possible worlds. The problem to be solved is the constructive formal representation of these concepts.

Constructive meaning of a sentence is usually defined by its verification conditions. If this is the case, a statement cannot be constructively justified if it cannot be verified. This does not mean that only actually verifiable sentences are constructively meaningful; constructive semantics is not tantamount to verificationism. If no conditions can be given under the fulfilment of which the sentence S would be verified, however, S is not constructively meaningful.

Talking about the empirical world, i.e., the world of sensory experience where factual statements apply, raises a particular constructive difficulty. In general, this empirical world is considered to be inexhaustible, i.e., there are infinitely many atomic facts to be learned about it. Most general statements concerning the empirical world are beyond definitive verification, including scientific generalizations. Under the meaning condition which is based on verifiability, these statements seem to be meaningless.

However, before going into discussion about constructive semantics for statements about the empirical world, it is in order to provide analysis of the empirical world. There is not just one possible empirical world – surely there are several ways things might be in the actual world. In philosophy, there is a special discipline called possible worlds semantics, which discusses what kinds of entity possible worlds are. For this study, it is enough to assume that possible worlds consist of all possible ways to describe states of affairs in the language at our disposal. The language in question will be that of first-order logic with identity and functions. In Rudolf Carnap's semantics (1937, 1946, 1947, 1962, 1952), linguistic representations of different possible states of affairs are called state descriptions (see section 2.1 below).

As mentioned above, the empirical world (and thus all the possible ways that

the empirical world might be) is infinite, not necessarily in the sense of being spatially infinitely extensive but inexhaustible. Assuming that there are parts of a possible world which transcend human cognitive capabilities, i.e., which are not knowable even in principle, is constructively problematic. In mathematics, this problem is overcome by appeal to potentially infinite structures, like the set of natural numbers, where the infinity of the structure is a consequence of a finitely cognizable rule. By knowing the structure of the set of natural numbers, one can always generate another natural number. There is no upper bound for the size of the set of natural numbers.

The properties of an infinite mathematical structure are coded in its definition. Let us consider the expression 'w is an empirical world'. It will be shown below that expressions in a finitary language (i.e., a language whose expressions are finite) do not suffice to represent the totality of infinite possible worlds even in the most elementary cases. In other words, the cardinality of the set of expressions in finitary language is less than that of the set of possible worlds.

All possible empirical worlds thus cannot be represented in a finitary language. On the other hand, it would be dramatic if adopting constructive semantics entailed that the possible empirical worlds must be finite simply because linguistic identification of some infinite possible worlds is not feasible. The constructively justified existence of an entity does not necessarily mean that all its features must be representable in a language. In this sense, constructive semantics is a way of thinking about the meaning of linguistic expressions, but it does not say whether something exists or does not exist outside language.

In constructive semantics one must be able to give the verification condition for the claim that 'w is an empirical world'. If w is infinite and cannot be grasped in its totality, there must be some other way of guaranteeing that it is a possible world; for example, from the way it is given to us.

Because there are insufficient expressions for all possible worlds, the statement 'w is an empirical world' is not expressible of all possible worlds. Hence, there seem to be possible worlds which cannot be denoted by a linguistic expression. On the other hand, if it is claimed that something exists, the assertion claiming this must have a verification condition in order to be meaningful.

Saying that an entity exists which cannot be talked about is a constructively meaningful assertion provided that there is a verification condition for it. Possible worlds, in the sense of alternative ways for the actual world to be, can also exist constructively even if there are not enough names for them, because the totality of possible worlds can be constructively given in such a way that it is clear that there are no names for all elements in the totality. This will become evident in what follows.

2.1 Cantor space

Let us start by assuming that the logically possible domains have denumerable cardinalities.[1] Hence, the space of possible worlds must include worlds with cardinality $|N| = \aleph_0$.

State descriptions in Rudolf Carnap's state description semantics are collections of atomic sentences and their negations. Every state description contains either the sentence itself or its negation for each atomic sentence of the language. For example, if the language contains the predicates $P(x)$ and $Q(x)$ and the individual constant a, the possible state descriptions of the language are

$$\{P(a), Q(a)\},\ \{P(a), \neg Q(a)\},\ \{\neg P(a), Q(a)\},\ \{\neg P(a), \neg Q(a)\}. \qquad (2.1)$$

The various state descriptions of the language are regarded as linguistic descriptions of possible states of affairs or possible worlds.

A Carnapian semantics contains at most a denumerable number of individual constants. Hence, the greatest cardinality among the domains of individuals of state descriptions is $\aleph_0$, i.e., that of N.

For the discussions in this study, it will be convenient to represent an infinite state description by an infinite binary sequence of 0:s and 1:s in which every atomic sentence is represented by a corresponding term. For example, if the language contains only one monadic primitive predicate $A(x)$, the atomic sentences corresponding to first three terms of the binary sequence would be $A(1), A(2)$ and $A(3)$. If the value of a term is 0, the corresponding atomic sentence is false; if the value is 1, the atomic sentence is true. This representation captures the idea of state descriptions as observation processes proceeding in time.

For first-order languages with more predicates and of higher arities, it is possible to use a set of binary sequences or even represent these predicates by a single binary sequence (see p. 51 below).

The (classical) set of binary sequences Ω is called the *Cantor space*. The constructive version of the Cantor space will be introduced in section 2.3 below.

However, although the Cantor space is said to consist of binary sequences representing all the logically possible state descriptions, this cannot literally be the case since there are not enough linguistic signs to exhaust the whole of the Cantor space, as shown below.

The set of binary sequences is nondenumerably infinite, i.e., it is not only infinite but has the cardinality of the continuum. For each term of the sequence, there are two possibilities, 0 and 1, which means that there are

$$2^{\aleph_0} \qquad (|N| = \aleph_0) \qquad (2.2)$$

[1] This assumption can be justified by the Löwenheim-Skolem theorem: all satisfiable sentences must be satisfiable in a denumerable domain.

possibilities of assigning truth values to atomic sentences in a countably infinite set of individual constants. Hence, the set of state descriptions must be uncountable, i.e., its cardinality is that of the continuum.

Consider an expression (e.g. a function $f : N \to \{0, 1\}$) assigning each term of the sequence a value 0 or 1) which denotes an infinite sequence of 0's and 1's. Since the set of expressions is denumerable, there can be only denumerably many state descriptions in the literal sense of the word. Hence, most of the elements in the Cantor space cannot be state descriptions in the literal sense of the word. This raises the question of whether the Cantor space of sequences is entitled to be called the space of state descriptions. After all, state description in Carnap's sense is a linguistic expression.

States of affairs can exist classically even if they are not denoted by linguistic expressions. The term "state description" can be used figuratively about states of affairs whose atomic parts can be described by individuals and predicates of the language in question. It is possible at least in principle to produce an answer to the question of whether an arbitrary atomic sentence is true or false in a given state description, although it is clear that no finite number of answers to such questions can distinguish a state description from an uncountable number of other state descriptions.[2]

2.2 Logical possibility

Can the continuum of state descriptions (see above) be dispensed with in formalizing the notion of logical possibility? In other words, is representing possible worlds by linguistic expressions enough for their adequate logical treatment after all?

Should this be the case, one would not have to define a notion of truth for a continuum of state descriptions, which in the case of constructive truth may prove to be impossible given the lack of sufficient expressions identifying the worlds. If one needs to define truth only for denumerably many state descriptions to capture logical possibility, the task of defining constructive truth becomes considerably easier.

Logical possibility plays an important role in inductive logic. It is essential to be able to represent all the logical possibilities for a sentence to be constructively true. There is no obvious connection between being able to represent logical possibilities and assigning probability measures to sentences, but such a connection will be established in chapter 3.

[2]A solution to linguistic representation problem is provided by Bricker (1987, p. 343) by adding an infinitary sentential conjunction to the language. However, such solutions are constructively problematic.

2.2.1 Re-interpretable names

The first attempt to do away with the continuum in defining logical possibility goes as follows.

Even though there are not enough names for all possible worlds, there might be a *possible* name for each world through the re-interpretation of names. Consider the expression w_a as referring to a certain possible world w'. In order to refix the reference, one would have to say that w_a refers to w'' instead of w'. However, for this the worlds must already have names (w' and w'' in this case), which leads to the recurrence of the original problem.

This also applies to valuation functions when they are interpreted as entities representable by linguistic expressions. Using just one name to refer to several worlds by changing the valuation function means, in fact, that one generates several names. For example, the valuations $v_1(w)$ and $v_2(w)$ can refer to two different worlds, but they are themselves two different expressions at the same time.

Hence, by reinterpreting the names one cannot increase the number of potential referents from $\aleph_0$ to a continuum.

2.2.2 Denumerable logical space

Another possibility is suggested by the following reasoning.

Suppose there is a meta-level language L^M consisting of expressions referring to state descriptions which, for all possible sentences, contains a single state description which makes the sentence true. L^M must be denumerable because there are only denumerably many sentences. One would not have to refer to the totality of logically possible worlds since one could represent the logical modalities by the expressions in L^M. A sentence is possible if it has a corresponding expression in L^M.

The first question that arises here is whether all the state descriptions required can be referred to by expressions. A sentence can be (classically) possible without an explicit reference to a truth-maker (i.e., a particular state description in which the sentence holds). It might be even the case that some truth-makers which are needed to provide the logical possibilities cannot be expressed linguistically. Hence, it is not clear without further investigations whether a set like L^M is feasible in classical semantics.

In Carnap's semantics, the sentence S is necessary, $N(S)$, if and only if it is logically true, i.e. holds in all possible state descriptions. S is possible, $\diamond(S)$ iff $\sim N(\sim S)$ is true (cf. Carnap 1947, pp. 174-175).

In classical semantics, this means that a sentence is possible if and only if it holds in at least one state description. In constructive semantics, $\sim N(\sim S)$ does not entail that S does this. The constructive meaning of $\sim N(\sim S)$ is that

the assumption that S is false in all state descriptions leads to contradiction, but $\sim N(\sim S)$ does not contain a reference to a particular state description where S verifiably holds.

Hence, $\diamond(S)$ can be constructively defined either as $\sim N(\sim S)$ or as saying that S holds at least in one state description.

Assume that the last one of these definitions is chosen for $\diamond$, i.e., $\diamond$ is defined by reference to a particular state description, and is thus a stronger notion of possibility than $\sim N(\sim S)$. It can then be proved that if $\sim N(\sim S)$ holds, $\sim \diamond(S)$ cannot hold, which can be seen below.

Observe that N and $\diamond$ can be treated like universal and existential quantifiers over the Cantor space of state descriptions. It is a theorem of intuitionistic predicate logic that

$$\sim (\exists x)P(x) \Leftrightarrow (\forall x) \sim P(x), \tag{2.3}$$

which translates into

$$\sim \diamond(S) \Leftrightarrow N(\sim S). \tag{2.4}$$

Assume that $\sim \diamond(S)$ holds. Then $N(\sim S)$ holds which, together with $\sim N(\sim S)$, yields a contradiction. Hence, $\sim \diamond(S)$ cannot hold.

This means that the assumption that there is no state description through which S can be proved leads to a contradiction with the assumption $\sim N(\sim S)$. Because of this, even if only the weaker notion of possibility $\sim N(\sim S)$ can be proved, there are consequences for the stronger notion with $\diamond$ as well. The assumption that there is no expression for a state description in which S is true would mean that $\diamond(S)$ cannot be constructively proved, i.e., $\sim \diamond(S)$ would be constructively true. But this is impossible if $\sim N(\sim S)$ is constructively true; hence the assumption that there is no expression for a state description in which S is true is contradictory. This means that even possibility in the weaker sense of $\sim N(\sim S)$ cannot do without an expression for such a state description where S is true.

The possibility of a sentence being claimed to be true in a state description without even the possibility of providing a denotation for the truth-maker is not constructive possibility. A sentence is constructively possible in state description semantics only if it can be proved in some state description, and it is not possible to prove this without being able to say which state description we are talking about. Since there can be only a denumerable number of identifiable state descriptions, all constructive logical possibilities can be represented by a denumerable number of identifiable state descriptions.

Now the question is how to select the relevant state descriptions to be included in L^M. This problem will be analysed in connection with the concept of extendible probability in chapter 5.

2.3 Constructive logical space

In constructive mathematics, the classical Cantor space can be replaced by a constructive *spread* of *lawless* binary sequences.

Lawless sequences are special cases of the more general concept of a choice sequence. Choice sequences – as formulated by Brouwer – consist of consecutive and infinitely proceeding choices of natural numbers by an idealized epistemic agent. This epistemic agent is sometimes referred to as an idealized mathematician who can complete any finite number of choices but is not capable of making an infinite number of choices. However, it is not possible here to dwell on the discussion concerning the ontological status of such an agent.

The agent chooses first number n_1, for example, then n_2 etc. without an upper bound. The *spread* is the totality of the sequences constructed in such a way, i.e., by choosing a number from the set of natural numbers at each stage.

Since choice sequences come into existence term by term in the course of time, the continuum of choice sequences exists only potentially, not actually. This is the essential difference between the spread of binary choice sequences and the classical interpretation of the Cantor space. In a sense there are non-denumerably many possibilities to implement a lawless sequence construction process, although no such possibility can ever actually be finished.

The properties of choice sequences can be determined either by knowing a rule that effectively outputs a value for a given position or knowing a finite approximation of the sequence. There is a one-to-one correspondence between the set N and the rule-determined or *lawlike* infinite sequence $\langle 0, 1, 2, ... \rangle$. The n'th term of the sequence can be computed by using the rule of adding one to the $n-1$'th term.

A lawless sequence is one whose terms are not governed by any restriction (other than the *a priori* restriction for the terms to be of the specified type, e.g. natural numbers); such a sequence is generated (and identified) by a process involving repeated arbitrary selection of one term after another. A partly free sequence is one in which some, but not total restriction may be imposed upon choices of terms. (Cf. Dummett 1977, p. 418, 423; Troelstra 1977, p. 12.)

The principle of open data states that the truth of any statement made about a lawless sequence can depend only upon some initial segment of it:

$$\varphi(\xi) \rightarrow \exists x \forall \eta (\overline{\xi} x = \overline{\eta} x \rightarrow \varphi(\eta)) \tag{2.5}$$

where $\overline{\xi} x = \langle \xi(0), \xi(1), ..., \xi(x-1) \rangle$, i.e., the initial segment of length x of ξ. In words: if φ holds for the lawless sequence ξ then there is an initial segment of ξ such that all lawless continuations of this sequence also satisfy φ (cf. van Dalen 1986, p. 313).[3]

[3]Limitation to lawless continuations is not essential — the quantification is just supposed to

The principle can be justified as follows: since $\varphi(\xi)$ is established after a finite number of values of ξ have been chosen, because at any time that is all the available information on ξ there is, the continuation of this particular initial segment is irrelevant; i.e., all continuations also have the property φ (op. cit., pp. 313-14).

One can use the same framework for analysing all types of choice sequence because they can be considered to represent different readings on the same scale. At one end of this scale there are the lawless sequences with no restrictions at all concerning the selection of terms, and at the other law-like sequences with total restrictions. One associates a spread law with each spread, which, when applied to any finite initial segment, determines whether or not the segment is admissible to the spread. At one extreme, the restriction set up by the spread law may be completely empty, which means that the sequence is a lawless one. At the other extreme, the restriction may fully determine the terms; in this case the result is a law-like sequence. (Cf. Dummett 1977, pp. 65-66, 423.)

2.4 Formalizing choice sequences

Let us consider the representation of choice sequences in a formal language. Would the formalization of choice sequences provide a solution to the problem of representing the totality of possible worlds?

In formal languages, variables usually range over sets or classes of objects. Permitting substitution of terms referring to choice sequences requires that the terms qualify as elements of sets over which the variables range.

Most constructive systems prescribe that an element of a set is effectively recognizable as being an element of that specific set. Let us see what this means for choice sequences.

In dealing with binary lawless sequences, one never knows anything more about such a sequence than what is given by a finite approximation like $0(1(0(0(...))))$. There can be only a denumerable number of such approximations, but it is usually held that the choice sequences themselves constitute a continuum.

How can one then be sure that the given expression $0(1(0(0(...))))$ really refers to an element of the set of binary sequences? One knows that the embodied initial segment is an initial segment of a binary sequence, but one is not in a position to recognize that the three dots in fact refer to a valid continuation of the sequence; one can only assume that they do so. In other words, one assumes that the three dots also refer to a binary sequence. But this means that one assumes that the object referred to by the dots is recognizable as being an element of the set of

apply to over lawless sequences.

binary sequences. This, in turn, entails that the object referred to by dots must be expressible by means of a finite expression since otherwise it would not be possible to recognize it to be an element or elements of the set of binary sequences.

The expression $0(1(0(0(...))))$ is thus effectively recognizable as being an element of the set of binary sequences only if "..." can be replaced by a finite expression of the language. This expression can be a variable with an open possibility of substituting finite expressions; this is the method used for representing choice sequences in the interpretation of Per Martin-Löf's non-standard type theory (1990) in standard constructive type theory. In other words, a choice sequence which transcends our cognition is represented by an infinite (but denumerable) number of cognizable objects. However, the question remains whether truth in infinity can be captured by such an interpretation of lawless sequences.

2.5 Observation sequences

In this study, the main emphasis is on sensory observations, which are considered to constitute the source of knowledge about the empirical world (i.e., the world outside the mind of the observer).

The focus of this study is thus not on the choices of an idealized mathematician. It will be argued below that the representation of sequences in formal semantics by means of approximations (see section 2.4 above) corresponds to the ontological status of observation sequences which consist of the observations of an *idealized observer*. The idealized observer is tentatively assumed to be capable of performing any finite number of observations but can never complete a series involving an infinite number of observations. Under this assumption it is feasible to consider a potentially infinite observation sequence. In other words, the observational setting itself could count as a verification for the infinity of the sequence. It will be discussed in section 2.5.3 below whether this analogy with choice sequences is justified.

2.5.1 The potential infinity of observation sequences

Postulating the existence of an infinite entity (e.g., a possible world) would require justification other than that which can be achieved by observation alone. It is clear that no finite number of observations can establish the existence of an infinite sequence of observations. Moreover, since observations take place in time, there can only be a finite number of observations at any point of time even in the classical sense (more about this in 2.5.2 below).

However, by analogy with lawless choice sequences, it is constructively possible to conceive of a lawless observation sequence which has no upper bound. In

observation sequences, the terms are chosen by "nature" instead of the idealized mathematician. This is different from lawless choice sequences. There the idealized mental agent has complete control over choosing the consecutive terms of the sequence, whereas in observation sequences the observer can merely observe what is being presented to her. The lawlessness of the sequence can be interpreted to mean that the working of the mechanism which generates the observations (i.e., "nature") is not known to the observer. It is not necessary to assume that the world outside the observer is non-deterministic.

The identity of the observation sequence is given by reference to a particular observational setting and to the point of time at which the observation sequence is initiated. For example, consider the observational setting of me tossing a coin. I start a new observation sequence today at 7.13 pm; this point of time and the rule which defines what will be observed (tosses of a particular coin) then define a specified observation sequence.

Observation sequences can, at least in their formal representation, be restricted by a spread law governing the future observations, which makes the sequences partly or completely deterministic.

The idealized mathematician of lawless choice sequences can decide at the outset that the process of selecting new terms in the sequences will never stop. Analogously, one can assume that the idealized observer of the observation sequence can construct an observational setting which will continue generating the observable phenomena infinitely. Consider, for instance, the coin-tossing example, where I can decide at the outset that I will continue tossing the coins, in the same way as I can decide that I will continue choosing new terms for my lawless choice sequence. (However, the soundness of the assumption that I can decide that my observations will continue forever is doubtful. The consequences of this fact will be discussed on page 31 below.)

In the coin-tossing example, the continuity of the sequence is under my control, but the outcomes of the tosses are not. If the outcomes were also under my direct control, the proper term for the sequence under consideration would be a lawless choice sequence.

2.5.2 Law-like sequences

This section discusses the question of whether it is possible to obtain knowledge about an infinite observation sequence by using mathematical induction. This, if possible at all, is certainly only possible for observation sequences which are law-like. If mathematical induction is a feasible method for law-like observation sequences, the fact that the sequence consists of observations does not alone make certain knowledge about its future behaviour impossible.

If one can prove properties for infinite sequences of observations in principle

by using mathematical induction, it makes constructive sense to say that such a sequence satisfies a sentence which is not finitely verifiable. This entails that finitely non-verifiable sentences could be constructively true in infinite sequences whose terms refer to observations (see below about the problems of this reference). This feature will prove important in explaining constructive truth and probability by means of extendible truth and probability in chapter 5.

One characteristic feature of lawless sequences (both choice sequences and observation sequences), is that observations and choices take place in time. This means that in the case of a series of observations, it is not possible to generate another choice or observation of the series in the same sense as it is always possible to generate another term of a law-like sequence, when considered purely as a mathematical object, by applying the law defining the sequence.

It is clear that observations are not instantaneous; they always take some period of time to make. Assume that making an observation will take a period of time denoted by δt. Assume, moreover, that at t_0 one has made n observations. It follows that there can be no more than n observations before $t_0 + \delta t$.

This reasoning applies to any given point of time, which entails that no more than a finite series of observations can exist at any given time. Assume that this series at t_0 consists of n first observations of the unlimited sequence α. This means that the question 'what is the m'th term $(m > n)$ in the sequence α?' has to wait until it can be answered.

On the other hand, one can never generate all the terms of any infinite sequence, not even if there is a deterministic function which could repeatedly be applied to generate consecutive terms. Any actually performed application of the function lasts for a non-negligible duration of time. Hence, most of the terms of a law-like sequence exist only potentially.

Consider now an observation sequence which is governed by a deterministic rule. The future observations belonging to such a sequence are fixed in advance. Despite the difference in the generation mechanism of terms between law-like observation sequences and law-like choice sequences, the truth conditions of sentences in both are essentially similar. This can be shown by examining the interpretation of the laws governing the terms of the sequences.

Consider the truth condition of the statement which expresses the governing law $w : N \to \{0, 1\}$:

$$\text{The value of the } n\text{'th term in the sequence } w \text{ is } w(n). \tag{2.6}$$

The constructive meaning of (2.6) is

$$\begin{aligned}&\text{It can be verified that the value of the } n\text{'th term}\\ &\text{in the sequence } w \text{ is } w(n).\end{aligned} \tag{2.7}$$

If (2.6) was interpreted as meaning that the value of the n'th term is $w(n)$ independently of any verification one would make a classical claim.

In law-like choice sequences, the law governing the sequence w yields a term value $w(n)$ whenever applied to a natural number n. The truth condition (2.7) thus holds by stipulation.

When the law represents a restriction on future observations, its interpretation is that every observed term will have the value determined by the function w. The method which is used to verify the values of certain terms in observation sequences is simply observing the terms. Hence, (2.7) means that it can be verified by observation that the value of n is $w(n)$, and this can take place only by waiting until n is observed. In other words, the law does not say that the value of n is $w(n)$ independently of any observation, but that there is a method to verify that n's value is $w(n)$, namely by waiting and observing n. The law both says that n will be observed and that by this observation its value can be verified to be $w(n)$.

To illustrate that the proof mechanism of mathematical induction is essentially the same irrespective of the interpretation of the law-like sequence as an observation sequence, consider the following proof by mathematical induction.

Suppose 1 in a binary sequence is interpreted to denote heads and 0 tails. Let us denote these by a monadic atomic predicate: heads is H and tails $\sim H$. Atomic facts about the terms are expressed by atomic sentences; they are verified by direct observation. The standard verification conditions of more complex sentences can be defined in the spirit of Heyting (1934).

Consider then the verification condition of a universal quantification: $(\forall x)P(x)$ is verified if there is a method of verifying $P(c)$ for an arbitrary c. In law-like sequences, the verification of P for an arbitrary c is based on mathematical induction which makes use of the law governing the sequence.

Now interpret x in $(\forall x)P(x)$ as ranging over the terms of some infinite law-like observation sequence w and $P(x)$ is a compound predicate which consists of occurrences of $H(x)$ and logical connectives. If $P(0)$ is true and if one can infer $P(x+1)$ from $P(x)$, by mathematical induction $P(c)$ holds for any c and thus $(\forall x)P(x)$ holds.

Although $P(x)$ cannot be directly verifiable by observation since it is a compound sentence, its verification condition is based on observing whether $H(x)$ holds or not. The inductive step in the mathematical induction involves an assumption concerning the verifiability of $P(x)$ which in turn involves an assumption concerning the observability of $H(x)$. Similarly, the truth of $P(x+1)$ means that it is verifiable on the basis of observing whether $H(x+1)$ holds or not. Thus the conclusion $P(c)$ means that P is verifiable for an arbitrary c on the basis of observing whether $H(c)$ holds or not.

Even if the observation sequence was not verifiably infinite, the law governing the sequence could imply that if there is another observation, it has a certain prop-

erty. In this case, mathematical induction could establish a theorem of the form 'if c is observed, then $P(c)$' for all c.

2.5.3 Mind-independent infinity

The domain of individuals must be defined somehow, i.e., what the objects one studies are must be stated. For instance, in the research process consisting of tossing a coin, the objects of study are the individual tosses.

The potential infinity of an empirical research process is thought to be a fact which is independent of the observer. In other words, it is not the observer who decides that the domain of individuals of the research is infinite. In the section above, the infinity of observation sequences was equated with the infinity of choice sequences in the sense that it can be concluded from the observational setting that will always be another term in the sequence.

However, the continuation of an observation sequence is not completely under the control of the observer, as contrasted to a choice sequence, which is completely controlled by the idealized mathematician. The observer may have an *intention* to toss a coin an indefinite number of times, but since nature is not under the observer's control, it cannot be known whether her effort will be successful. Assuming that the observation process is under the control of the observer would amount to assuming that the observer is omnipotent concerning a certain region of space-time. While it is certainly conceivable that an idealized mathematician, deprived of all human characteristics except existence in time and space, is in control of her own choices, it is not equally plausible that an idealized observer is in control of the mechanism which produces the observations.

In the coin-tossing process, the domain of individuals, consisting of consecutive tosses, is generated by the observer. However, in many cases the observer does not generate the individuals only makes observations about them. Consider the countable infinity of a constellation of stars (see, e.g., Fletcher 2002). If the constellation of stars is taken as the domain of individuals, the proper constructive semantics cannot be based on knowing that the domain is infinite, because this cannot be verified by observation. Hence, it is even less justified to talk about infinite observation sequences when the domain is not generated by the observer than in the coin-tossing case.

2.5.4 Coordinate languages

Carnap introduces the term *coordinate language,* where the individual terms do not designate objects but simply positions about which it is not stated whether

there are objects to be found (1937, p. 141; 1947, p. 75; 1962, p. 62).[4] The positions are denoted by expressions like o for the first position, o' for the successor position, then o'' and so on.

It is constructively legitimate to adopt coordinate languages because one can always generate a successor position by adding $'$ to the expression referring to the concerned position.

However, interpreting the individual terms in a coordinate language as referring to the empirical world is more problematic. For example, space-time cannot be judged to be infinite by a simple linguistic stipulation. In other words, to use a coordinate language with infinitely many positions to represent the coordinates in physical space-time, one must assume that space-time is indeed infinite.

If the coordinate language contains infinitely many positions, it is not knowable whether every position will correspond to an observation. Atomic sentences which contain individual terms not referring to observations do not describe states of affairs "out there" .

2.5.5 Possibly infinite observation sequences

Instead of being potentially infinite, observation sequences should rather be characterized as *possibly* infinite, in the sense that according to the knowledge of the observer, the assumption that the sequence will not stop does not lead to a contradiction.

If it is not knowable that the observation sequence is infinite, a formal structure which knowably has no upper limit would have no justification in representing the sequence. It is thus appropriate to look for a representation of observation sequences which is not based on the notion of potential infinity. A formalization which does not commit one to infinite state descriptions would obviously need to be based on expressions referring to finite objects, as suggested in section 2.4 above for other reasons, those pertaining to the constructive definition for an element of a set. This kind of formalization can be achieved by means of Martin-Löf's nonstandard type theory, for instance (cf. Holm 2003; Ranta 1992). However, the notions of truth and probability can be discussed without a type-theoretical interpretation of observation sequences. This topic will be elaborated in 5.1.

[4]The physical space-time coordinates or temporal positions are given as examples of such a language in Carnap (1937, p. 141) and Carnap (1962, p. 62), respectively.

Chapter 3

Constructive epistemic probability

This chapter discusses various challenges and difficulties in defining a constructive notion of epistemic probability, in which probabilities are interpreted as justified or rational degrees of belief. The question of whether the concept of infinity provided by sequences in a Cantor space is appropriate for a constructive formalization of observation sequences is put aside for the moment.

3.1 The axioms of probability calculus

The probability calculus is conventionally defined using the following formal machinery, in which the space of elementary events is denoted by Ω and the event space is some class F of subsets of Ω.

Consider an arbitrary non-empty space Ω. A class F of subsets of Ω is an *algebra* if it contains Ω itself and is closed under the formation of complements and finite unions (cf. Billingsley 1995, p. 19-20):

(i) $\Omega \in F$,
(ii) $A \in F$ implies $A^c \in F$,
(iii) $A, B \in F$ implies $A \cup B \in F$.

A σ-algebra is also closed under the formation of countable unions:

(iv) $A_1, A_2, ... \in F$ implies $A_1 \cup A_2 \cup ... \in F$.

Condition (iv) naturally implies (iii).

One can assign probability measures to algebras (cf. Billingsley 1995, p. 22), but they are usually defined on σ-algebras.

Any probability measure $P : F \rightarrow R$, where R is the set of real numbers, is usually required to fulfil the axioms introduced by Kolmogorov (1933):

(i) $P(A) \geq 0$ for all $A \in F$,
(ii) $P(\Omega) = 1$,
(iii) If $A_i \in F (i = 1, 2, ...)$ and $A_i \cap A_j = \emptyset \quad (i \neq j)$,
then $P(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} P(A_k)$.

Axiom (iii) is called countable additivity or σ-additivity. Sometimes only a more restricted condition of finite additivity is required:

(iv) If $A, B \in F$ and $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$.

The probability axioms do not provide an *interpretation* of the probability function, i.e., they do not tell how the values of P are calculated. The axioms only outline the limiting conditions for the possible interpretations.

In constructive semantics, one must interpret axioms (i)-(iii) in a constructive fashion. The truth requirement of the axioms means thus that they must be verifiable in the proposed constructive semantics. However, it may be questioned whether the axioms should also hold for a constructive probability function or whether they should be amended somehow.[1]

A brief discussion concerning this issue is provided in section 3.2.4, where one immediate argument against the axioms is refuted.

When it comes to distinguishing between the traditional interpretations of probability, this study is focused on the epistemic interpretation, in which probability is formulated in terms of the degrees of rational belief of an epistemic agent. A constructive interpretation for epistemic probabilities is sought for, with an accompanying argument that the proper interpretation for an epistemic probability function must be constructive.

Carnap's state description semantics, which in its original form is classical logical semantics, provides a logical framework for discussing the meaning of probability. Carnap's explication is perhaps the best-known logical interpretation of probability.

Logical interpretation of probability is a branch of epistemic probabilities. In the logical interpretation, probabilities are assigned to propositions (or sentences of first-order logic in Carnap's case). The state descriptions represent the space of logical possibilities. Assigning probability values to sentences by means of logical relations derived in the space of state descriptions is motivated by the presumption that one can arrive at objective or *a priori* probabilities in this way, thus representing the objectively rational degrees of belief. For example, if a sentence is true in half of the state descriptions, it can be assigned the probability $\frac{1}{2}$. On the

[1]Should this be the case, another approach would of course be to declare that constructive probability is not possible since no such function can satisfy the axioms of probability.

other hand, by giving unequal weights to state descriptions, the probability value can also be something else in this case.

However, when one moves to considering the infinite logical space, problems begin to occur. Carnap does not thoroughly relate the asymptotic limit interpretation (see section 4.1) of probability functions in an infinite domain to the definition of truth in an infinite state description, which consists of an infinite class of atomic sentences and their negations. The asymptotic limit definition of probability deals only with the finite domains of individuals.

Moreover, the concept of *constructive probability* is a difficult one from the outset. There is no interpretation of constructive probability which is based on the notion of the constructive truth of a sentence in an infinite state description (or more generally, in an infinite possible world or model). The problem is that truth as provability or verifiability has not seemed a useful concept in this context because observation is usually not enough to prove general statements about the empirical world. Those possible worlds whose formalization has made them accessible to formal proof (for example, worlds whose atomic facts are given by computable functions) occupy such a small part of the probability in infinite domains of individuals (in fact, infinitesimally small) that provability in these worlds does not, at the outset, seem very significant with respect to probability considerations.

The above issues will be discussed in section 4. The chapter begins with a treatment of inductive logic as an explication of epistemic probability (section 3.2). In section 3.2.3, it is suggested that the constructive interpretation of inductive logic is the most credible one, if inductive logic is to be an explication of epistemic probability. Section 3.2.4 discusses the relation between the betting interpretation of probability and constructive semantics.

3.2 Inductive logic and epistemic probability

Inductive logic or logical probability in the sense of Carnap (1962) could be conceived of as a formal discipline in its own right, without being considered as an interpretation (or to use more precise language, explication) of the more or less vaguely defined concept of probability. In this view, inductive logic simply provides logico-mathematical methods of establishing the degree of entailment or, using Carnap's terminology, degree of confirmation between evidence and hypothesis (the function to express this will be called the entailment function below, denoted by Ent), but this degree is not claimed to be an explication of the concept of probability. The degree of entailment would simply express the (weighted) proportion of state descriptions in which the hypothesis is true of those state descriptions where the evidence is true.

However, the reason for inductive logic is its use as the foundation of probability. Inductive logic is often referred to as a logical interpretation of probability or logical probability.

The epistemic motivation of probability in general is to provide degrees of belief under uncertain circumstances, i.e., where deductive methods of obtaining knowledge are not available.

Suppose that the following is considered to hold: from the fact that the degree of entailment between h and e is p it follows that if somebody knows e and nothing else, this person is justified in believing in h to the degree p and also acting according to this belief. This view is put forward for example in Carnap (1962), p. 44. (Note that e can also be a logical truth.) Observe the words "nothing else" here, which are required by the principle of total evidence (cf. Carnap 1962, p. 211). If one also believes something other than e which affects the probability of h, one is certainly not justified in believing in h to the degree p.

On this account, inductive logic is regarded as a foundation for rational inductive inferences in the sense that it provides a method for determining justified (or rational) degrees of belief.

In what follows, the circumstances under which leap from the mathematical degree of entailment to justified degrees of belief is possible will be examined in more detail.

3.2.1 Probability in infinity as verifiability

In this section some preliminary remarks will be made concerning the constructive interpretation of Carnap's inductive logic and constructive probability in general.

In inductive logic, probabilities are established on the basis of the truth and falsity of sentences in all the alternative possible worlds or state descriptions. For the probability value of a sentence to be defined, it is not necessary to be able to decide whether the sentence is true or false in the actual world, of which it is not known which formal state description it corresponds to; it is only required that the truth value of the sentence can be determined in the state descriptions belonging to the logical space.

In constructive semantics, a sentence is true or false in a state description of the logical space only if it is verifiably true or verifiably false.

Constructive logical probability as probability which is based on a constructive notion of truth is determined by the (weighted) proportion of those state descriptions in which the sentence is verifiable. The question of interpreting the probability function classically or constructively is independent of the interpretation of truth. However, it is hardly possible to adopt constructive semantics for truth in the object language and classical semantics for the probability function. For example, in the Carnapian asymptotic limit approach to probability, the existence of a limit

is expressed by an existential quantifier. It is not a consistent semantic stance to interpret the quantifier classically if truth is otherwise interpreted constructively.

A question arises, however, when the meaning of constructive logical probability is compared to that of classical probability. The meaning of the classical probability of S includes that it is the probability with which the actual world is such that S is true. This is the reason for using probabilities in decision-making. Constructively speaking, however, the actual world may be such that S is true in it only if S is verifiable in it. The question is whether the actual world *is* one of those state descriptions where S is verifiable – and the 'is' in italics must here be interpreted constructively.

The problem is that there is no link from the constructive degree of entailment of S to the probability of the actual world turning out such that S is verified. Even if a certain ratio p of the state descriptions satisfy S constructively, this does not mean that S is verifiable in the actual world with the probability p because it may be the case that which state description corresponds to the actual world can never be found out.

Classically speaking, the actual world may correspond to some state description where S is constructively true even if this is not known. In fact, S would then be classically true in the actual world. In classical semantics, the link from the degree of entailment of S to the probability of the actual world being such that S is true in it does exist. It thus seems that classical probability is a better guide than probability defined as verifiability in the actual world, but then one must of course accept classical truth for S. The probability of classical truth provides a different kind of rational guide than the probability of constructive truth. The probability of constructive truth is about coming to know S, whereas the probability of classical truth has no relation to the state of knowledge of the cognitive agent.

The epistemic motivation of probability is the cases in which no certain knowledge is available. Such cases include those in which the sentence cannot necessarily be verified or falsified in a finite time. If probability of $(\forall x)P(x)$ is about the actual world turning out such that the S is true in it, then it is clear that $(\forall x)P(x)$ must have zero probability. However, this limitation on the use of probability is quite drastic, since it can no longer be used as epistemic guidance for situations which go beyond finite knowledge. Probability is reduced to situations in which uncertainty currently prevails but which will eventually be decided. In section 4.1.3, it will be shown that this kind of notion of constructive probability can lead to absurdity.

On the other hand, one may ask what kind of rational guide to life would such probability be where the probability of S is not about the actual world being such that S is true in it. Since the most obvious constructive interpretation of "being such" – namely, verifiability – does not work, one must come up with some other constructive interpretation for the phrase or, alternatively, adhere to

classical probability.

Even if S is not known to be true in the actual world, it may not be known to be false either. In the absence of any knowledge, one has to act on the basis of probabilities. In this case the probability is not about S being true in the world, but is rather about the world being such that $\sim S$ will not be proved when new data is discovered. Although this kind of non-falsifiability is not an adequate basis for probability, as will be shown by the remark concerning (4.12) on p. 52 below, it is a step toward achieving a formulation of probability based on constructive notions.

Since defining probability in terms of knowability or verifiability in the actual world is not a reasonable goal at all, the requirement that S should be verifiable in those state descriptions which are considered as positive outcomes for S becomes obsolete. If probability of S cannot, in any event, be interpreted as the probability of true verifiability of S (i.e., verifiability in the actual world), there is little need to define probability on the basis of verifiability in the alternative state descriptions.

This does not mean that S's is verifiability in a given state description w has no significance for rational decision-making. The point is that the significant effect can be attained with a less stringent observable property than the verifiability of S in w; namely, with a property which is close to the non-falsifiability of S in w, as mentioned above. The positive outcomes of S with respect to probability are to be constructively defined, but not as verifiability of the truth of S. This situation will be elaborated in chapter 5, where a constructive formalization of probability will be suggested.

3.2.2 Knowable degree of belief

A justified degree of belief does not have to be *actually* maintained by any particular cognitive agent. However, only such a term can refer to a justified degree of belief which could, at least in principle, denote somebody's actual degree of belief. Hence, a justified degree of belief must in principle be knowable, otherwise it could not be anyone's degree of belief.

This means that in Carnap's view, which was cited above on page 36, the cognitive agent must also know, beside e, that the degree of entailment between e and h is p.

A justified degree of belief is obtained by using sound methods. The degree of entailment is considered to be such a method. If the degree of entailment is knowable, then a method is known which, when implemented, will yield the value of the degree of entailment. By applying this method, one thus gets to know the degree of entailment.

If the degree of belief is to be determined by the degree of entailment, knowing the latter is a precondition for knowing the former. Hence, the only situations

in which some degree of belief can be known are those in which the degree of entailment is known or a method of obtaining it is known.

Degree of entailment is thus useful for determining degree of belief only if it is knowable. This already suggests that perhaps only constructively justified degrees of entailment, in which the entailment relation is interpreted according to constructive semantics, are useful or even sound in determining justified degrees of belief. The issue will be examined more closely in the next section.

If this view is accepted, statements of the form

$$Ent(h \mid e) = p, \tag{3.1}$$

where Ent is a function from a pair of sentences to a value in the range $[0, 1]$, are true only if they are verifiable; Ent is interpreted as saying that e entails h with the degree p. Furthermore, if follows from the constructive meaning explanations for the quantifiers going back to Heyting (1934) that the existentially quantified statement

$$(\exists x)[Ent(h \mid e) = x], \tag{3.2}$$

saying that there is a degree of entailment between e and h, means that (3.1) is verifiable for some particular natural number p.

The existential quantification comes to play an important role when dealing with the asymptotic limit interpretation of Ent in infinite domains because asymptotic limit probabilities are not always decidable (see, e.g., Grove, Halpern & Koller 1996). One can classically justify statements of the form

$$(\exists x)(LimEnt(h \mid e) = x) \tag{3.3}$$

without being able to specify a value for the limit, while in the constructive approach the limit exists only if it is knowable. If (3.3) is interpreted constructively, it follows that

$$(\exists x)(\text{the justified degree of belief=x}), \tag{3.4}$$

but not if (3.3) is interpreted classically. Hence, the question arises whether one should adopt constructive semantics in dealing with justified degrees of belief.

3.2.3 Constructive probability and justified degrees of belief

Whether justified degrees of belief suggested a constructive interpretation of probability, was discussed in the previous section, although no compelling argument was offered. Most naturally constructive probability would also mean a constructive interpretation of truth for the sentences of the object language. Talking about

constructive probability (the provable probability statements) of classical truth is certainly an odd, if not inconsistent combination. However, if the reason for adopting constructive probability is simply the fact that it suits with degrees of belief best, there is no obligation to adopt a more general constructive philosophy or semantics (concerning the sentences of the object language, for example).

The classical probability of a sentence can exist even where the probability is not known. If the classical probability of h is not known, then the justified degree of belief in h is not known. In this case, there is no justified degree of belief of h because a degree of belief must be known, but one could still consider the classical probability of h as being interpreted as the justified degree of belief of h *whenever* this probability is known. The formalized probabilities would then be considered as limiting conditions which the justified degree of belief has to fulfil. If a degree of belief is to be justified, it must abide by the values given by the probability function, but the function does not necessarily effectively yield the value in every situation.

However, if there are probabilities which cannot be known, these probabilities cannot be interpreted as rational degrees of belief. Maintaining that there are classical probabilities is tantamount to explaining the notion of probability as something other than a rational degree of belief.

Even if epistemic probabilities should be understood constructively, however, must the notion of truth for the sentences of the object language be constructive as well?

To analyse this question, consider the usual formula for conditional probabilities:

$$P(h|e) = \frac{P(h\&e)}{P(e)} \tag{3.5}$$

If truth is interpreted classically, this definition says that h's classical truth has a certain probability when e is classically true. To set the justified degree of belief according to (3.5), e must also be all the relevant evidence known to the agent (cf. the principle of total evidence on p. 36 above).

To adjust the justified degree of belief one thus requires known evidence, not simply classically true evidence. If e is known and nothing else which is relevant is known, one can set the justified degree of belief in h according to (3.5).

Consider now the following. If e is known, the actual world must correspond to a state description in which e is classically true, but then even more is true about the actual world than that e is classically true – namely, that e is known to be classically true.

If e is a finitely decidable sentence, this distinction does not matter. Whenever a decidable sentence is classically true, it can also be known to be true. However, if e is not a decidable sentence, its knowability does not coincide with its classical

truth, in which case, sentence e can be classically true even if it is not knowably true.

Recall once again the principle of total evidence: in determining the justified degree of belief, one needs to take all the available evidence into account. If e is known to be true, it is not simply classically true, but also knowably true. Hence, the probability of h should not be updated with the probability of e's classical truth, but with that of e's knowable truth. If one uses the probability of e's classical truth, one does not take into account the fact that e is knowably true and thus violates the principle of total evidence.

This suggests that to be consistent with this principle, one should apply constructive semantics in the object language when discussing the probability of truth of h and e. Probabilities would thus be probabilities of constructive truth.

Gilbert Harman has discussed another line of reasoning which, he claims, both leads to constructive semantics in the object language and calls into question the validity of a particular theorem of probability calculus (see section 3.2.4 below).

3.2.4 Harman and betting interpretation

Harman (1983) raises a doubt concerning the validity of

$$P(A) + P(\sim A) = 1. \tag{3.6}$$

Harman formulates his point in terms of the subjective or betting interpretation of probability, which is a branch of epistemic interpretation of probabilities in terms of degrees of belief. In the betting interpretation, the probability of a proposition is assumed to represent the odds one would require before betting for or against the truth of the proposition.

Many betting quotients can be rational, but some are not. Some betting quotients are not coherent in the sense that the bettor cannot possibly win. In the betting interpretation, it is argued that a probability assessment cannot be rational if it cannot be interpreted in terms of a coherent betting quotient.

For Carnap in 1962, logical probability meant a fair betting quotient (cf. Carnap 1962, pp. 165-167, 237). A logical probability can always be rephrased in terms of a betting quotient. This is a stricter requirement than mere coherence: Carnap seems to advocate the view that there is one rational betting quotient, which is provided by degree of confirmation. Nevertheless, the point here is that according to Carnap, logical probability must also have a formulation in terms of betting quotients.

Harman (1983, p. 242) argues that the betting interpretation of probability in fact implies constructive semantics in the object language (Harman uses the term intuitionism) since what settles a bet is not the truth or falsity of the proposition

in question, but the *discovery* that the proposition is true or the discovery that it is false. Since the basic doctrine of constructive semantics is precisely that truth and falsity mean the same as discovery of truth and discovery of falsity, the betting interpretation should adopt constructive semantics.

The fact that it cannot in general either be discovered that some fact obtains or that it does not obtain means, according to Harman, that one should assign a positive probability to the possibility that the issue is never settled. This implies, he claims, that theorem (3.6), which is usually derivable (the preconditions for its derivability will be discussed below) from the standard axioms of probability, is not valid for subjective probabilities.

Harman does not, however, explicitly point out the fact that (3.6) is usually considered to be derivable from the probability axioms; he merely refers to Field (1977) for the assumption that any reasonable subjective conditional probability function must satisfy it. The derivability from probability axioms makes the case more serious. One must ask whether constructive semantics is in fact incompatible with standard probability axioms.

Let us look more closely at how theorem (3.6) is derivable from the probability axioms.

It follows from the additivity axiom that

$$P(A \vee \sim A) = P(A) + P(\sim A). \tag{3.7}$$

Let us assume that $\sim A$ can be interpreted as being the complement of A, i.e., that $\sim A$ covers the cases (elementary events / state descriptions) where A does not hold:

$$\sim A = \Omega \setminus A. \tag{3.8}$$

Then either A or $\sim A$ always hold and thus

$$P(A \vee \sim A) = P(\Omega) = 1. \tag{3.9}$$

Result (3.6) follows from (3.7) and (3.9).

The crucial assumption in the proof is (3.8), which is not an axiom of the probability calculus. If this assumption is not accepted, (3.6) cannot be derived.

Harman (1983) can be interpreted as saying that assumption (3.8) cannot be accepted in the subjective interpretation of probability, i.e., that there are sometimes cases (elementary events / state descriptions) in which the truth or falsity of the sentence will not be discovered, meaning that there are cases which belong neither to A nor to its complement.

However, Harman's argument is not formulated in an entirely consistent manner. For an event or proposition to have a certain probability in the betting interpretation, it would have to be discoverable. This is also Harman's point: the

possibility of a proposition never being discovered must have some probability. However, in the general case it is not possible to discover in a finite time that something will not be discovered in a finite time. Therefore, if the probability of a proposition is to be considered in terms of discovering it, it is not possible to assign a probability value to *not* discovering that A or $\sim A$.

Despite this problem, Harman's point about the non-validity of (3.6) in the constructive interpretation stands. One does not even have to assume as much as Harman does. One can dismiss (3.8) by simply stating that nothing guarantees that it is true. There are state descriptions in which A cannot be proved, i.e., $\sim A$ holds, but these do not necessarily encompass all the state descriptions where A has not been proved. It may be the case that neither A nor $\sim A$ is known to be true.

Hence, because (3.8) and thus (3.9) are not constructively valid, the constructive interpretation of probability does not necessarily contradict the standard axioms of probability. Expression (3.6) is simply not derivable from these axioms. In other words, subjective probability does not lead to adjustments of probability axioms.

Chapter 4

Formalizing probability in infinity

This chapter considers the problem of constructing a probability measure in the infinite space of infinite state descriptions.

The notion of asymptotic probability in Carnap's inductive logic is an explication of the concept of a justified degree of belief in a formal language of infinitely many individual constants (and thus also infinitely many infinite state descriptions). However, it will be argued that Carnap's approach seems to lack a proper connection with truth in infinity.

Instead of using the asymptotic approach, probability can be defined for the infinite logical space, which would establish a direct connection between truth and probability in infinity. This chapter concludes with a discussion concerning probability measures in a Cantor space formalisation of state descriptions. Although a probability measure based on verifiable truth in infinity turns out to be problematic, a probability measure for classical truth in infinity can be constructed. It will be shown that this measure cannot represent the probability of constructive truth in infinity.

4.1 Infinite domains according to Carnap

How do we assign probabilities to members of the nondenumerably infinite logical space of infinite state descriptions? This problem is tackled by Carnap (1962, p. 289) using an asymptotic limit procedure. The probability of a sentence in infinity is defined as the asymptotic limit of its probability in consecutive finite domains $D(n)$ containing the first n individual constants (or, following Carnap, systems of language L_n):

$$P_\infty(S) = \lim_{n\to\infty} P_n(S). \tag{4.1}$$

On this definition, the probability values are calculated on the basis of truth in finite state descriptions, not infinite ones. The limit is the limit of the probability that the sentence can achieve in arbitrarily large finite domains. However, one should also establish a connection with the truth of the sentence in infinite state descriptions. This relation is not established by Carnap nor by subsequent commentators on his works.

Let us first discuss the case of sentences without quantifiers. This case is simple, since such sentences can be discussed in the framework of propositional logic. The sentences in propositional logic are truth-functional, i.e., their truth value can be determined on the basis of the truth value assignment to the atomic sentences occurring in them. This means that if a sentence without quantifiers is true in a certain state description (which corresponds to a truth value assignment to the atomic sentences), it is true in all the finite and infinite extensions of this state description.

It follows from this that the meaning of a quantifier-free sentence is the same in different domains of individuals (or language-systems in Carnap's terminology) since its truth-conditions remain the same.[1] The probability of such a sentence is also the same in every finite and infinite domain where it is a sentence at all.

The case of quantified sentences is different. Carnap explains the meanings of quantifiers in finite language systems by means of conjunction and disjunction over all the individual constants of the language (Carnap 1962, p. 60, 62). It follows that their meanings are language dependent, because the stock of individual constants varies between the systems L_n (it is, however, assumed that the domains of individual constants are nested, i.e., $D(i) \subset D(i+1)$ for all i). The universal sentence $(\forall x)P(x)$ in infinity is equivalent to the infinite class of instances of $P(x)$. The meanings of $(\forall x)P(x)$ in the consecutive systems L_n with finite n do nevertheless "converge, so to speak, toward its meaning in L_∞". (Carnap 1962, p. 60.)

However, it is not clear what Carnap means by convergence of meaning here. According to Carnap, the range $R_n(S)$ of the sentence S is the class of state descriptions of L_n in which S is true. The rules of ranges determine the range of any sentence in L_n.

The idea of a rule of range is to point out state descriptions in which the sentence in question holds. If one knows how to do this, if one knows the rule of range, one knows the meaning of a sentence in the Wittgensteinian sense (cf. Carnap 1947, p. 10).

The rules of range specify the meaning of a sentence to the extent that it can be expressed without specifying the meaning of each atomic sentence, i.e. without

[1] Constructively one would have to speak about verification conditions, but that does not make any difference in this case.

specifying interpretation of its predicate symbols and individual constants. The whole meaning of a sentence also includes these and is expressed by its rule of truth.

In predicate calculus with identity (which the Carnapian systems L_n do contain, see Carnap 1962, p. 61), there are sentences which are not finitely satisfiable but are valid in the infinite domain. In consecutive finite domains, the ranges of such sentences are empty, and thus in no way converge to the range in the infinite domain. The idea of convergence presumes that these sentences are excluded from the discussion.

Are the ranges of $(\forall x)P(x)$ in consecutive finite domains approximations of the range of $(\forall x)P(x)$ in the infinite domain? The range of $(\forall x)P(x)$ in L_n contains state descriptions which have extensions in larger domains that do not satisfy $(\forall x)P(x)$. Since an infinite number of extensions of members of this range will be excluded from subsequent ranges of $(\forall x)P(x)$, every rule of range which selects a number of finite state descriptions to constitute the finite range of $(\forall x)P(x)$ also selects an infinite number of possible extensions which do not satisfy $(\forall x)P(x)$ in L_∞. In any L_n the rule of range only points out state descriptions in which $(\forall x)P(x)$ *can* possibly hold in larger domains of individuals, these being those that belong to the range of $(\forall x)P(x)$ in L_n. The rule of range in the infinite domain would be a method of pointing out the state descriptions in the infinite domain where $(\forall x)P(x)$ holds. The question now is whether the repeated applications of the rules of range in the finite domains somehow converge toward a rule of range in the infinite domain.

Consider the rule of range for $(\forall x)P(x)$ in L_∞, which says that $P(a_n)$ must hold for all a_n when n is arbitrarily large. This means that no finite L_n can comprise anything except an infinitesimal proportion of all the individual constants for which $P(x)$ must hold. Hence, the rules of range in L_n are not approximations of the rule of range in L_∞. Even if a given state description $(\forall x)P(x)$ fulfils its rule of range in a given state description of L_n, it is not closer to truth in L_∞ since it would still have to fulfil the rule of range in an infinite number of larger domains. To put this in another way, establishing the truth of $P(x)$ for a certain finite set of instances does not bring us any closer to establishing the truth of $P(x)$ for an infinite number of instances.

Although the asymptotic probability of $(\forall x)P(x)$ is associated with the rules of range in consecutive finite domains, no repeated application of a rule of range for a finite domain of individuals can point out state descriptions in the infinite domain in which $(\forall x)P(x)$ holds. The best that is achieved is ruling out state descriptions in which $(\forall x)P(x)$ does not hold. This, however, is insufficient for a meaning explanation.

Note that it is certainly true that the proportion of the range of $(\forall x)P(x)$ of all finite state descriptions tends to zero, i.e., its probability value tends to zero,

but this is not the same as the convergence of finite meanings toward the infinite meaning. It might be said that the probability values of $(\forall x)P(x)$ in consecutive finite domains do converge toward the probability of $(\forall x)P(x)$ in the infinite domain, but then one must define the probability of $(\forall x)P(x)$ in infinity in terms of its meaning in infinity.

The conclusion from the discussion above is that it is not clear how the meaning of $(\forall x)P(x)$ in finite domains converges toward its meaning in the infinite domain. One remaining task in the programme of inductive logic is thus to find a concept of truth in infinity which would match the asymptotic limit approach to probabilities. If this is not possible, one must conclude that the Carnapian probability of a sentence S is something other than the probability of S's truth in an infinite domain.

The problems discussed in this section necessitate a brief digression to other approaches to probabilities in infinity than the asymptotic one adopted in Carnap (1962).[2] An alternative approach which uses infinite domains instead of a series of finite ones provides a more straightforward connection between the notions of truth and probability in infinity. However, such an approach will be shown to be constructively problematic.

4.1.1 An infinite number of state descriptions

The probability space of state descriptions can be represented by means of the Cantor space Ω, which is a collection of infinite binary sequences, as explained in section 2.1 above. Such sequences can be considered as representing infinitary state descriptions.

In order to assign a probability to a state description in the Cantor space, one should specify the state description under consideration. An enumeration of its elements, i.e., a list of binary digits (or truth values of atomic sentences), will not do for this purpose since no list can be infinite. Only finite parts of such infinite worlds can be specified by this kind of enumeration.

It is possible, however, to denote complete state descriptions with functions over individual constants. The combination of these functions would then determine one state description.

However, as was seen in section 2.1, there are not enough expressions to exhaust the logical space. There is only a denumerable number of expressions, but the whole Cantor space has a non-denumerable cardinality. This means that the set of those state descriptions that can actually be denoted by a linguistic expression is an infinitesimal minority of all state descriptions and it is clear that one

[2] See also the discussion in Carnap 1962, p. 303, touching the question of a probability measure in an infinite set of state descriptions.

cannot found a semantics of probability on such a minority, at least not without further arguments.

4.1.2 Integer chosen at random

De Finetti discusses the example of choosing a positive integer at random (De Finetti 1972, p. 86). While the probability of a particular integer being selected is zero, the whole space of positive integers has the probability 1. Hence, countable additivity does not hold because the sum of probabilities of integers is zero and the probability of the union of integers is 1.

In De Finetti's example, there are an infinite number of elementary events which all have an equal probability. In such cases, the probability measure cannot be countably additive. For example, if one takes the state descriptions as elementary events, their union must have a probability of 1 although the sum of their individual probabilities is zero. However, for reasons mentioned in section 4.1.1, infinite state descriptions will not be considered as elementary events in what follows below.

4.1.3 The neighbourhood approach

Although the Cantor space is originally a classical notion, it can be given a constructive meaning as a spread of lawless binary choice sequences (see section 2.3 above). The binary digits 0 and 1 of the sequences may be interpreted as the truth values false and true respectively. It is obvious that the binary sequences can be interpreted to correspond with state descriptions of a language of predicate calculus with one monadic predicate.

However, sections 4.1.3.3-4.1.3.6 below apply only to genuine Cantor spaces whose sequences cannot be interpreted by means of finite sequences. Moreover, most of what appears in this chapter is valid in classical set theory only.

The purpose of this discussion is first to demonstrate how one could construe a probability measure for sentences in the infinite domain and why this definition would fail in the constructive setting (in section 4.1.3.6). Second, the discussion points out the fact that the difficulties of defining the concept of probability in infinity are characteristic of constructive semantics. Third, because at least some classical probability measure for state description semantics can be formulated by using subsets of the Cantor space, the asymptotic limit approach to probabilities (like in Carnap 1962) is actually needed more in a constructive version of state description semantics (with a constructive concept of truth) than with the classical one.

4.1.3.1 Neighbourhoods

Instead of complete state descriptions of the Cantor space, it is possible to assign a probability measure to certain sets of state descriptions, which are called *neighbourhoods*. The neighbourhood approach could be employed to solve the problem of reference since there is an expression to denote each neighbourhood – the set of neighbourhoods is thus denumerable, although the neighbourhoods themselves are non-denumerable.

The *initial segment* of length j of a sequence w in the Cantor space is provided by the *restriction function*

$$r_j(w) \in W(j) \quad (w \in \Omega), \tag{4.2}$$

where $W(j)$ is the set of sequences of length j. The neighbourhoods of the Cantor space are sets of infinite sequences which have a common initial segment.

The neighbourhood defined by the initial segment w_j is the set

$$J(w_j) = \{w \in \Omega | r_j(w) = w_j\}. \tag{4.3}$$

Neighbourhoods allow us to assign a probability measure to sets of infinite sequences. One can define the *Lebesgue measure* for neighbourhoods along the lines of Martin-Löf (1968, p. 91). The Lebesgue measure P^L gives the probability value 2^{-n} to each neighbourhood defined by means of an initial segment of the length n. For example, the whole Cantor space has the probability value of $2^0 = 1$ and a neighbourhood which is defined by one binary digit has a probability of $\frac{1}{2}$. The probability of the neighbourhood approaches 0 when the length of the defining initial segment grows without an upper limit.[3]

4.1.3.2 Truth in a neighbourhood

What is the relation between the concept of truth and the probability measure which is defined over neighbourhoods?

Truth in a neighbourhood $J(w_j)$ can be defined as truth in all the state descriptions. This can be formally presented as follows:

$$J(w_j) \models S =_{Df} (\forall w \in J(w_j))(w \vDash S). \tag{4.4}$$

According to this definition of truth in a neighbourhood, S is true in an infinite state description w only if it is true in every infinite state description belonging to $J(w_j)$. What is common to all these state descriptions is the initial segment w_j.

[3]The Lebesgue measure corresponds to Carnap's $m^\dagger$ measure, which assigns each state description equal probability.

The first observation is that the concept of truth in a neighbourhood results in a restricted applicability of probabilities. To consider $J(w_j)$ as a positive outcome for the probability of the sentence, the sentence ought to be true in all the state descriptions belonging to it. Hence only sentences can have positive probability which are verifiable on the basis of some finite body of evidence. Because non-valid sentences which contain universal quantifiers do not fall into this group, such sentences must receive zero probability value.

One of the main uses of probability is to evaluate the degree of certainty of sentences that cannot be established from a finite number of observations. Scientific hypotheses are in general established only with some probability, not verified. For example, even though it can never be concluded on the basis of observations that all ravens are black, this should not be enough to conclude that the probability of the proposition is zero. Scientific hypotheses should be enabled to have a non-zero probability; if this is not the case, they cannot not be confirmed by empirical evidence at all.

The second observation is about the sentence

$$(\exists x)(\forall y)C(x,y). \tag{4.5}$$

This sentence is neither verifiable nor falsifiable. Hence, neither (4.5) or its negation are verifiable on the basis of any finite initial segment, which means that, on the view presented above, both should be impossible and thus have zero probability.

Before further deliberations on (4.5), it must be shown that a language containing a two-place relation can indeed be interpreted in the Cantor space.

The following logical space corresponding to a two-place relation can be used for encoding the example sentences above.

Let

$$\mathbf{C} = \{\Omega\}_{k=1}^{\infty} \tag{4.6}$$

be a countable ordered set of Cantor spaces, corresponding to one two-place relation. Let us denote the x'th Cantor space by Ω_x.

The truth value of an atomic sentence $C(x,y)$ is determined by the y'th digit in the sequences belonging to Ω_x. For example, the first Ω in (4.6) consists of binary digits corresponding to the atomic sentences $C(1,1), C(1,2), ...$, the second Ω of binary digits corresponding to the atomic sentences $C(2,1), C(2,2), ...$ and so on.

One can also encode all the state descriptions into a single Cantor space Ω in the following manner:[4] the first binary digits represent the truth values of the

[4]In considering languages of many predicates of different arities, it might be more convenient to give up the space of binary sequences and resort instead to sequences with more options at each node. Each node would then represent a complete state description.

atomic sentences

$$C(1,1), C(1,2), C(2,2), C(1,3), C(2,3), C(3,3), \\ C(1,4), C(2,4), C(3,4), C(4,4)... \tag{4.7}$$

Consider the sentence $(\forall y)C(a,y)$ for some constant a. Since there is only one state description which satisfies this sentence, the Lebesgue measure of the sentence in the uncountable Cantor space must be zero (how to construe the measure for this sentence will be explained starting from section 4.1.3.3 below). Moreover, the union of state descriptions which satisfy $(\exists x)(\forall y)C(x,y)$ is clearly the union of state descriptions satisfying $(\forall y)C(a,y)$ when a is consecutively replaced by each individual constant of the language. This union is thus a countable set with a zero Lebesgue measure. It is thus reasonable to expect that $\sim (\exists x)(\forall y)C(x,y)$ has a probability of 1 or at least a positive probability.

Moreover, an adequate concept of probability distinguishes between a sentence and its contradiction on the basis of received evidence. It follows that even a non-verifiable sentence like $\sim (\exists x)(\forall y)C(x,y)$ should have a positive probability. Since it is not reasonable to expect that all non-verifiable sentences have zero probability, one cannot argue that scientific hypotheses should have zero probability simply because they are non-verifiable. Should one defend such a claim, one should also accept that sometimes both a sentence and its contradiction must have zero probability.

The zero probability of $(\exists x)(\forall y)C(x,y)$ and its negation seems also to be in contradiction with the theorem

$$P(S \vee \sim S) = 1. \tag{4.8}$$

However, $\sim S$ is not the complement of S with respect to truth in a neighbourhood. The sentence $\sim S$ means that S's falsity can be established on the basis of some initial segment. This is not always the case when the truth of S cannot be established on the basis of the initial segment.

Let us define this concept of complement more precisely.

Let $\mathbf{S}_\Omega$ denote the set of state descriptions in Ω where S can be verified on the basis of their initial segments. Let

$$\mathbf{S}_\Omega^{\mathbf{C}} = \Omega \setminus \mathbf{S}_\Omega = \{\mathbf{w} \in \Omega | \mathbf{w} \notin \mathbf{S}_\Omega\}. \tag{4.9}$$

In other words, if S cannot be verified on the basis of an initial segment of w, then $w \in \mathbf{S}_\Omega^{\mathbf{C}}$. This in turn means that $\sim S$ must be classically true in at least one state description in $J(r_j(w))$ for every j.

Let us denote the negation of S corresponding to this complement by $\neg S$. The truth condition for this negation is thus: $\neg S$ is true in the neighbourhood $J(w_j)$ iff $\sim S$ is true in at least one state description in $J(w_j)$.

This concept of negation has the consequence that

$$\neg(\exists x)(\forall y)C(x,y) \tag{4.10}$$

is true in every neighbourhood because $(\exists x)(\forall y)C(x,y)$ cannot be verified on the basis of any initial segment. Hence,

$$P((\exists x)(\forall y)C(x,y) \vee \neg(\exists x)(\forall y)C(x,y)) = 1 \tag{4.11}$$

and thus no contradiction with $P(S \vee \neg S) = 1$ arises.

Should one thus replace $\sim$ by the complement negation $\neg$? The problem with $\neg$ is that the the truth value of $\neg S$ can change with new evidence. For example, sometimes $\neg(\exists x)P(x)$ can be true until evidence verifies $(\exists x)P(x)$. This is a very unintuitive conception of negation and does not reflect the notion of falsity in a state description. Moreover, it follows that $\neg(\exists x)P(x)$ can receive a probability of 1 in a particular neighbourhood but a probability of 0 if more evidence becomes available. This is not a reasonable concept of probability.

To sum up the discussion here, a concept of truth cannot reasonably be founded on verifiability on the basis of initial segments.[5] This also applies to the concept of probability.

But what if one relaxes the requirement that the sentence must be true in all state descriptions of the neighbourhood? Perhaps one could only require that the sentence be true in *some* state description of the neighbourhood, as formalized below.

$$J(w_j) \models' S =_{Df} (\exists w \in J(w_j))(w \vDash S) \tag{4.12}$$

Consider now the earlier example of (4.5) and its negation. Neither is falsifiable by any finite initial segment, which means that each of them is true in some state description of every neighbourhood. This means in turn that both must be assigned a probability of 1, which is obviously absurd.

4.1.3.3 An algebra of neighbourhoods and unions of neighbourhoods

A way to assign probabilities to (monadic) first-order sentences will now be outlined. The basic idea, present at least in Jeffrey (1971), is to define the propositions (i.e., the set of infinite sequences) corresponding to the sentences by means of countable unions and intersections of neighbourhoods and to assign probabilities to propositions formed in this way. Since there are only denumerably many sentences, one can form a proposition for each sentence.

[5]This result may cause problems to concepts like Crispin Wright's superassertibility (1992), according to which a predicate is superassertible if it is assertible in some state of information and then remains so no matter how that state of information is enlarged upon or improved.

The set-theoretical operations of union and intersection can be defined constructively (cf. Bishop 1967, p. 64). However, it is not claimed that the discussion below can be carried out completely in a constructive fashion. The aim of the treatment here is to describe one possible way to assign probabilities to sentences, following the idea of the range of a sentence consisting of all the state descriptions in which it is true. Only after presenting the proposal it is possible to discuss (in section 4.1.3.6) where the problems of this proposal with respect to constructive semantics are located. These problems are considered to establish that the notion of constructive truth of S in an infinite lawless sequence which is verifiable only by looking at finite initial segments of the sequence cannot be the basis of constructive probability of S in the Cantor space representing infinite state descriptions. At the same time, the discussion relates the goals of this study to abstract probability theory. Chapter 5 will present another, hopefully more successful attempt to define the notion of constructive truth and the corresponding notion of constructive probability.

The first observation in the treatment by means of unions and intersections of neighbourhoods is that the set of neighbourhoods does not fulfil the usual requirement of a set of events since it is not closed under set-theoretical operations. The union of two neighbourhoods not having common elements (e.g., the union of two neighbourhoods defined by the binary initial segments 000 and 111) is not necessarily a neighbourhood. Hence, one needs to generate a set of events which is closed under finite unions.

Consider Ω to be the set of sequences and F to be the class of neighbourhoods and finite unions of neighbourhoods. It is easy to see that F is an algebra:

(i) $\Omega \in F$ holds trivially;
(ii) If $A \in F$, $A^c = \Omega \setminus A$ is clearly a neighbourhood or a finite union of neighbourhoods;
(iii) If $A, B \in F$, $A \cup B$ is clearly a finite union of neighbourhoods.

Consider the above Lebesgue measure P^L of neighbourhoods (cf. p. 49), which can be defined in an algebra which also includes unions of neighbourhoods. That this is the case can be seen from the following. Suppose that A and B are neighbourhoods. One of the following three options then holds: $A \subseteq B$, $B \subseteq A$ or $A \cap B = \emptyset$. If $A \cap B = \emptyset$, $P^L(A) = \frac{1}{2^x}$ and $P^L(B) = \frac{1}{2^y}$, then $P^L(A \cup B) = \frac{1}{2^x} + \frac{1}{2^y}$ for some x and y.

4.1.3.4 The countable additivity of P^L

Let us then see whether P^L fulfils the condition of countable additivity (see p. 34) usually required from probability measures. This will turn out to be an important property of P^L.

The countable additivity condition also concerns infinite countable unions which belong to F. Hence, we must first see whether if there are infinite countable unions in F.

Suppose first that

$$\bigcup_{k=1}^{\infty} A_k \in F \tag{4.13}$$

where the A_k are disjoint neighbourhoods. The union $\bigcup_{k=1}^{\infty} A_k$ must clearly consist of neighbourhoods of diminishing size when the index k grows since the probability of any neighbourhood is positive.

It follows from (4.13) that there is a finite union of disjoint neighbourhoods B_l

$$\bigcup_{l=1}^{n} B_l = \bigcup_{k=1}^{\infty} A_k \tag{4.14}$$

because F is a class of neighbourhoods and finite unions of neighbourhoods.

Moreover, it is easily shown that

$$(\forall k)(\exists l)(A_k \subseteq B_l). \tag{4.15}$$

since it is clear that for a given A_k, $A_k \cup B_l \neq B_l$ cannot be the case for all B_l because otherwise A_k would contain elements which do not belong to $\bigcup_{l=1}^{n} B_l$ (two neighbourhoods cannot overlap without one being a subset of the other).

It is evident that the following holds for all A_i which are subsets of a particular B_l:

$$\bigcup_{i} A_i = B_l. \tag{4.16}$$

If B_l contains sequences which do not belong to $\bigcup_{i} A_i$, these sequences must belong to some A_k which is not a subset of B_l. However, according to (4.15), this A_k must be subset of another member of $\bigcup_{l=1}^{n} B_l$, but this is impossible because the members of $\bigcup_{l=1}^{n} B_l$ are disjoint.

From (4.16) and (4.15) we get:

$$\text{Every set in } \bigcup_{k=1}^{\infty} A_k \text{ belongs to a union } \bigcup_{i} A_i \text{ which equals some } B_l. \tag{4.17}$$

Since the sets in $\bigcup_{k=1}^{\infty} A_k$ are non-empty, disjoint and infinite in number, not all neighbourhoods B_l can each be covered by a finite union of such A_k's, otherwise

these finite unions would cover the whole finite union $\bigcup_{l=1}^{n} B_l$, which cannot be the case. Hence, at least some $B'_l \in \bigcup_{l=1}^{n} B_l$ must equal an infinite union of A_k's.

Every neighbourhood naturally contains an infinite number of smaller neighbourhoods. Moreover, every neighbourhood contains a finite number of smaller neighbourhoods with initial segments of a given length. Consider then all neighbourhoods included in B'_l which have the initial segment of length m. The union of these neighbourhoods equals B'_l itself.

Suppose now that each of these neighbourhoods is a subset of 1) some A_k in the infinite union which equals B'_l or 2) some finite union of these A_k's. Suppose first that 1) is the case. It follows that there must be a finite number of A_k's which cover B'_l. But this cannot be the case since B'_l equals an infinite sequence of disjoint A_k's. Suppose then 2). It follows again that there is a finite number of A_k's which cover B'_l. Contradiction.

Hence, there is at least one neighbourhood in B'_l (say, $J(w'_m)$) which is not a subset of one of the A_k's or a finite union of A_k's. It follows that there must be at each level $m+i$ some extension w'_{m+i} of w'_m such that $J(w'_{m+i})$ is not included in any A_k; if this was not the case, i.e., if all neighbourhoods of the form $J(w'_{m+i})$ for some i were included in some A_k, $J(w'_m)$ would itself be included in that A_k, which is not the case. Moreover, if $J(w'_{m+i})$ is not included in any A_k, the same argument applies to it, and so on.

An infinite series of consecutive extensions of w'_m thus defines an infinite series of neighbourhoods which are not included in any A_k. But is this infinite sequence w'_∞ of extensions of w'_m itself included in some A_k, despite the fact that none of the approximating neighbourhoods is?

Since $w'_\infty \in J(w'_m)$ and $J(w'_m) \in B'_l$, it holds that $w'_\infty \in B'_l$. Therefore, w'_∞ belongs to an infinite union of A_k's. Because this union consists of disjoint sets, w'_∞ must belong to one of the A_k's in the union.

Suppose now that w'_∞ is included in one of the A_k's, say $A_{k'}$, which is a neighbourhood. Then for some j, it holds for the initial segment w'_{m+j} of w'_∞ that $J(w'_{m+j}) \subseteq A_{k'}$. Contradiction. Therefore, w'_∞ does not belong to the infinite union of A_k's. Contradiction. It follows that assumption (4.13) is not consistent, i.e., that an infinite countable union of disjoint non-empty sets cannot be included in the algebra F. The countability condition thus holds trivially for P^L in F.

4.1.3.5 A probability measure for sentences

Since the event space consisting of the set of neighbourhoods and unions of neighbourhoods of the Cantor space is an algebra and P^L is a countably additive probability measure, by a well-known theorem there is a unique probability measure of the σ-algebra generated by the set of neighbourhoods and unions of neighbourhoods which coincides with P^L (cf. Billingsley 1995, p. 36-37).

In the new event space, one can form events corresponding to quantified sentences. In other words, one can define sets of state descriptions (binary sequences) which are ranges of quantified sentences.

For example, in a language with one monadic predicate $H(x)$ representing the case that the toss denoted by x is heads, the universal quantification $(\forall x)H(x)$ can be represented as follows.

First write

$$V_{H(x)}(k) = \{w \in \Omega | H(k) \text{ is true in } r_k(w))\}. \tag{4.18}$$

The sentence $(\forall x)H(x)$ corresponds to

$$\bigcap_{k=1}^{\infty} V_{H(x)}(k). \tag{4.19}$$

In other words, (4.19) is the definition of the range of $(\forall x)H(x)$, $R_\infty((\forall x)H(x))$.

It is clear that (4.18) is a union of disjoint neighbourhoods with all possible initial segments of length k, where w_k is a sequence of 1's of length k; hence, (4.19) is an intersection between neighbourhoods.

The sentence $(\forall x)H(x)$ is thus interpreted as a set of sequences (namely, of one sequence 1111... when '$H(x)$ is true' is represented by the x'th term value 1).

The existential quantification $(\exists x)H(x)$ corresponds to the set of sequences

$$\bigcup_{k=1}^{\infty} V_{H(x)}(k), \tag{4.20}$$

which is again obtained from neighbourhoods by combining them with operations on sets.

The presentation above applies only to monadic events. It will now be shown that the case of one two-place predicate can also be represented in a similar fashion, but with a slightly more complicated formalism.

Recall the coding of a single dyadic predicate into a single Cantor space in (4.7). Let us denote by $\mathbf{V} : N \times \Omega \to \{0, 1\}$ the function which denotes the binary value of a given term in a given sequence of Ω.

The sentence $(\forall y)C(1, y)$ corresponds to the set

$$\bigcap_{k=0}^{\infty} \{w | V(1 + \sum_{i=0}^{k} i, w) = 1\}. \tag{4.21}$$

Each element in the set $\{w|V(1 + \sum_{i=0}^{k} i, w) = 1\}$ of the above expression is a sequence in Ω whose $1 + \sum_{i=0}^{k} i$'th term is 1. Consider all initial segments of

such sequences. Each of the initial segments defines a neighbourhood. The union of these neighbourhoods equals $\{w|V(1+\sum_{i=0}^{k} i, w) = 1\}$. Hence, (4.21) is an intersection between unions of neighbourhoods and thus belongs to the σ-algebra in question.

The general case is $(\forall y)C(a, y)$, where a is an arbitrary individual constant:

$$\bigcap_{k=0}^{\infty}\{w|V(\sum_{i=1}^{a} i + \sum_{i=0}^{k} i) = 1\}. \tag{4.22}$$

The range of the sentence $(\exists x)(\forall y)C(x, y)$, $R_\infty((\exists x)(\forall y)C(x, y))$ is thus

$$\bigcup_{a=1}^{\infty}\bigcap_{k=0}^{\infty}\{w|V(\sum_{i=1}^{a} i + \sum_{i=0}^{k} i) = 1\}. \tag{4.23}$$

The sentence $\sim (\exists x)(\forall y)C(x, y)$ is simply defined as the complement of this set. The complement operation $\mathbf{C}$ can be defined as follows:

$$\mathbf{C}(A_\Omega) = \Omega \setminus A = \{w \in \Omega | w \notin A\}. \tag{4.24}$$

4.1.3.6 Elements of constructive sets

It has been shown above that one can at least classically assign probabilities to the sentences $(\exists x)(\forall y)C(x, y)$ and $\sim (\exists x)(\forall y)C(x, y)$ in the Cantor space of infinite binary sequences or state descriptions. However, there is a problem when one tries a constructive interpretation of the ranges of sentences as defined above.

It was shown in the preceding section how one can build $R_\infty(S)$ for some quantified sentences in predicate calculus with one monadic predicate as well as in a language with a single two-place predicate. It remains to show that the formulated sets really contain state descriptions where the corresponding sentence is true.

One should thus prove that if S is true in w, $w \in R_\infty(S)$ and vice versa. It turns out to be impossible to prove the following: if $w \in R_\infty(\sim (\exists x)(\forall y)C(x, y))$ *and* if $R_\infty(\sim (\exists x)(\forall y)C(x, y))$ is uncountable as one expects it should be, then $\sim (\exists x)(\forall y)C(x, y)$ is *constructively* true in w.

The statement $w \in R_\infty(S)$ means constructively that it can be proved that w is an element of $R_\infty(S)$. This can be done either by appealing to the rule that defines w or, where w is a lawless sequence, by showing that the neighbourhood defined by some initial segment $r_k(w)$ belongs to $R_\infty(S)$.

Consider the range

$$R_\infty((\exists x)(\forall y)C(x, y)), \tag{4.25}$$

which consists of a countable number of state descriptions (cf. p. 51 above). The range

$$R_\infty(\sim (\exists x)(\forall y)C(x,y)), \tag{4.26}$$

which was defined above to be the complement of (4.25), should thus consist of an uncountable number of state descriptions. Since there can be only countably many rule-defined sequences, the only possibility for this is that there are uncountably many lawless sequences in (4.26).

Consider the lawless state description α. To know that α is in (4.26), one should know that $(\exists x)(\forall y)C(x,y)$ cannot be true in α. It is possible to know this only on the basis of an initial segment $r_i(\alpha)$ for some i. But then one should know that $\sim (\exists x)(\forall y)C(x,y)$ is true in $J(r_i(\alpha))$, which is impossible since it is not verifiable on the basis of an initial segment.

Hence, no sequence which is not given by a rule can be proved to belong to (4.26). It follows that the range (4.26) cannot correspond to constructive truth.

This entails that construing the Lebesgue measurable sets on the basis of neighbourhoods along the lines above is not easily paired with any notion of constructive truth. Sentences which ought to have a positive measure, but are not verifiable or falsifiable form the hub of the problem. Their positive probability in an infinite uncountable set of state descriptions cannot be interpreted as the probability of constructive truth in a state description since no reasonable concept of truth for lawless sequences exists. Therefore, a constructive probability based on the notion of constructive truth will be attempted taking a different approach in Chapter 5. This approach will be based on the probability of an initial segment being a part of an infinitely proceeding rule-governed sequence satisfying S.

Chapter 5

Extendible probability

5.1 Introduction

This chapter offers a formalization of constructive probability in infinity. The new concept is referred to as *extendible probability,* which will be based on a new formalization of constructive truth in infinity, referred to hereafter as *extendible truth.*

As was seen above (cf. Ch. 3), the difficulty in the finitistic asymptotic limit approach is the connection to truth in infinity, in that the limit does not designate the joint probability of infinite state descriptions satisfying the sentence.

In order to define the constructive probability of truth in infinity, one should indicate what kinds of entity the elementary events (i.e., the possible worlds or state descriptions) are. State descriptions as observation sequences were discussed in section 2.5 above.

In section 2.5.3, some doubt was cast on the infinite character of observation sequences. Their infinity is better characterized as possible infinity, not in the sense that is usually meant by potential infinity but rather that it is not known whether the observation sequence will stop or not.

Truth in this kind of possibly infinite sequence cannot be based on the assumption that the sequence will not stop. However, one cannot expect that the sequence will stop either. Hence, the state descriptions as representations of observation sequences and constructive truth in state descriptions must be defined such that the epistemic nature of observation sequences is taken into account.

First it will be established (cf. section 5.2 below) that in predicate calculus without identity, a sentence true in a state description is also constructively satisfiable in any larger cardinality. Hence, in predicate calculus without identity, probability in a finite domain stands for the joint probability of those finite state descriptions in which the sentence is true and which can be extended to any

greater cardinality (including $\aleph_0$) so that the sentence is true in all the extensions. It follows that every finite and satisfiable extension of these extensions can also be extended to a larger cardinality, and so on. Therefore, if a sentence S is true in w_i, w_i has consecutive extensions up to infinity which satisfy S. In the calculus without identity, the asymptotic limit probability stands for the limit of the joint probability of such state descriptions.

This idea needs some further refinement to become a full-fledged formulation of the probability of truth in possible infinite observation sequences. A precise notion of extendible truth based on consecutive extensions of a state description will be introduced in section 5.4.

However, it holds that the so-called axioms of infinity like (5.16) below cannot be true in a finite domain, which means that their probability is not captured by the asymptotic approach. For these reasons, the notion of extendible probability captures the infinite only for sentences that are satisfiable in finite domains. This might be considered as a disadvantage of the notion. On the other hand, notions of constructive truth in the spread of infinite sequences do not give rise to a reasonable notion of probability, as shown above in Chapter 3.

The extendibility property will be proved for predicate calculus without identity. The significance of the proof is that it shows that the concepts of extendible truth and corresponding extendible probability to be introduced below have an important application in predicate calculus.

5.2 Satisfiability in greater cardinalities

Leblanc proves that every sentence in a *model set* for a given language L of predicate calculus without the identity symbol '$=$' or function symbols is true in some state description of L (1983, pp. 219-22). [1]

What is the significance of Leblanc's proof? Model sets, introduced by Hintikka (1955), are consistent sets of sentences fulfilling certain conditions. A *term extension* of L is any language that is exactly like L except for having countably many terms apart from those of L (cf. Leblanc 1983, p. 195.) It follows from the definition that a model set for L is a model set for any *term extension* L^+ of L.

According to the definition in Hintikka (1969, pp. 57-58), the set of sentences true in a state description is a model set.[2] Hence, if a sentence S is true in a state description w_k of L_k, it belongs to a model set of L_k and thereby to a model set of L^+. By Leblanc's proof, S is then true in a state description of L^+. (Note that the union of S and the atomic sentences of w_k is true in w_k and thus in a state

[1] Instead of using the term state description, Leblanc talks about truth value assignments for the atomic sentences of L.

[2] Hintikka calls such sets extended model sets.

description of L^+). In other words, if S is true in a state description of L_k, it is true in a state description of the term extension of L_k; and this is precisely what was to be proved.

Let us now examine more in detail the part of this proof which is described in Leblanc (1983).

The T_1 rewrite $T_1(S)$ of S is defined as the result of substituting the alphabetically earliest individual constant in place of each constant of S which does not occur in M. The *truth value associate* α_M of a model set M of L^+ is a complete truth value assignment to the atomic sentences of L^+ and thus analogous to a state description. (Cf. Leblanc 1983, pp. 219-220.)

Since Leblanc's own proof for

$$\alpha_M(A) = \alpha_M(T_1(A)). \tag{5.1}$$

in Leblanc (1983, p. 220) is fairly concise, I will reformulate it in a more detailed manner.

For each atomic sentence A of L^+, $\alpha_M(A) = T$ if $T_1(A)$ belongs to M and F otherwise. Consider first the case $A \in M$. Then $T_1(A) = A$ and thus $T_1(A) \in M$. It follows that $\alpha_M(A) = T$ and $\alpha_M(T_1(A)) = T$. Therefore, $\alpha_M(A) = \alpha_M(T_1(A))$. Let us then consider the case $A \notin M$. Assume first that $T_1(A) \in M$, which means that $\alpha_M(A) = T$. In this case, since $T_1(A) \in M$ and thus $T_1(T_1(A))$ in M (because $T_1(A)$ and $T_1(T_1(A))$ are the same), $\alpha_M(T_1(A)) = T$ holds and hence $\alpha_M(A) = \alpha_M(T_1(A))$. Assume then that $T_1(A) \notin M$. It follows that $\alpha_M(A) = F$. Because $T_1(T_1(A)) \notin M$, $\alpha_M(T_1(A)) = F$ holds as well. Hence also in this case and thus always, $\alpha_M(A) = \alpha_M(T_1(A))$.

The next step is to prove that

$$\alpha_M(P(a_1/x)) = \alpha_M(P(a_j/x)) \tag{5.2}$$

for an arbitrary j, if all the individual constants in $P(a_1/x)$ occur in M.

Let $P'(a_j/x)$ for any j be a substitution instance of $P(a_j/x)$ in which all quantified parts are replaced by their substitution instances; $P'(a_j/x)$ is thus a quantifier-free sentence of L^+.

$T_1(P'(a_j/x))$ is a truth-function of atomic sentences, being the sentence which is obtained when each atomic sentence A_P in $P'(a_j/x)$ is replaced by $T_1(A_P)$. Consider an arbitrary atomic sentence $T_1(A_P)$ in $T_1(P'(a_j/x))$. Since by (5.1), $\alpha_M(A_P) = \alpha_M(T_1(A_P))$ holds, every $T_1(A_P)$ in $T_1(P'(a_j/x))$ can be replaced by A_P without altering the truth value of $T_1(P'(a_j/x))$ in α_M. Hence, $\alpha_M(T_1(P'(a_j/x))) = \alpha_M(P('(a_j/x))$. Since the truth value of a quantified sentence depends only on the truth values of its substitution instances, it follows that

$$\alpha_M(T_1(P(a_j/x))) = \alpha_M(P(a_j/x). \tag{5.3}$$

Observe that $T_1(P(a_j/x)) = P(a_1/x)$ because all the individual constants of $P(a_1/x)$ occur in M and thus the only one in $P(a_j/x)$ which possibly does not is a_j. By (5.3), $\alpha_M(P(a_j/x) = \alpha_M(P(a_1/x)$. QED.

Following the technique adopted in Leblanc (1983, p. 221), which proceeds by induction on the length of sentences, it is easy to show using the above result that every sentence in M is true in α_M. The interesting cases are those with quantifiers. Consider, e.g., the sentence $(\forall x)P(x)$ which belongs to M. Then, by definition of a model set, $P(a_i/x)$ belongs to M for each a_i that belongs to M. By the induction hypothesis, $P(a_i/x)$ is true in α_M for each such a_i. Hence, by (5.2), $P(a_j)$ is true in α_M for each a_j which does not belong to M and thus $(\forall x)P(x)$ is true in α_M. The case with the existential quantifier is similar.

According to the above result, in predicate calculus without identity, the probability of S in each finite domain in fact denotes the joint probability of finite state descriptions which are initial segments of infinite state descriptions satisfying S. The asymptotic limit of such finite probability is then considered to be the probability of S in infinity.

What connects this concept of probability to constructive truth in infinity? After all, in each finite initial segment w_n contributing to the probability of S, it is only required that S be true in w_n and that S be *possible* in infinity; truth in infinity is a different issue. This topic will be discussed in what follows.

5.3 Infinity represented by finite structures

According to the approach of Carnap (1962), the elementary events of the probability space defined by the language L_n of the cardinality n are finite state descriptions of the elements in the domain of n individuals $D(n)$.

Events, in general, are sets of elementary events; in this case, they are sets of state descriptions. The range of a sentence is such an event. A finite state description w_n, belongs to the range of the sentence S if S is true in w_n. Carnap's approach operates at the level of finitary events without a definition of an infinite probability space.

However, one does not need infinite domains of individuals (infinite in the sense of $\aleph_0$) for representing an infinitely proceeding observation sequence. The concept of extendible truth in consecutive finite domains will be used to explicate truth in an infinitely proceeding observation sequence.

5.3.1 Nested domains

The concept of extendible truth was outlined above informally in 5.1. This concept, beside establishing a connection with truth in infinity, will also solve another

problem which arises from the original Carnapian limit definition of probability.

Some preliminary explanations concerning the formalism are in place before turning to consider more carefully the relation between the asymptotic limit probability and truth.

Carnap's inductive logic has only one domain (i.e., set of individual constants) for each cardinality and, in addition, the domains are nested in such a way that

$$D(i) \subset D(j), \tag{5.4}$$

when $j > i$, where $D(i)$ and $D(j)$ denote domains with the first i and j individual constants respectively. Individual constants are thus linearly ordered in the meta-language; Carnap uses the expressions $a_1, a_2, ..., a_n$ to denote the first n individual constants. (Cf. Carnap 1962, p. 58.)

Carnap (1962) does not himself provide an explicit argument for the necessity of this condition. However, it is needed in the limit approach for the following reasons.

First, it is clear that a domain has to be associated with each natural number greater than some k, i.e., $D(i)$ must denote a domain for each $i \geq k$. It is also clear that domains can have no upper bound in size, which means that there can be no i' such that

$$(\forall j)(D(i') \geq D(j)). \tag{5.5}$$

Consider then a sentence with the occurrence of some individual constant a. Well-definedness of the limit probability (4.1) requires that

$$(\exists m)(\forall n \geq m)(a \in D(n)). \tag{5.6}$$

The nestedness condition (5.4) guarantees that (5.6) is fulfilled and, moreover, it seems to be the simplest condition applying to domains of individuals which does this.

More justification for the nestedness assumption can be derived from the idea of possibly infinite cardinality. If domains could not be ordered like this, a series of them could not be considered as a domain without an upper bound, which is needed for representing observation sequences (see below). Hence, nestedness is necessary for the notion of extendible truth to be introduced below.

Since the nestedness condition imposes a linear order on the individual constants, one can replace the constants $a_1, a_2, ..., a_n$ by the natural numbers $1, 2, ..., n$, the advantage of this being that predicates concerning the order of the individual constants are immediately possible in the object language.

5.3.2 Observation sequences and sequences of possible observations

How does an infinite sequence of nested finite domains differ from an infinite domain? Both notions seemingly refer to infinity. The assumption that there are infinitely many nested finite domains means that there is an infinite number of individuals in the union of these domains. Infinity of individuals is thus imported into the semantics even without assuming an infinite domain of individuals. Are we then not committed to infinite structures, which was considered problematic in the conclusion of section 2.5.3 above?

In what follows it will be argued that the semantics of consecutive finite domains is the most appropriate one for constructive modelling of a possibly infinite observation sequence, i.e., where it is not known whether the process will continue without an upper bound (cf. 2.5.3). The individuals of this semantics are interpreted as observations in the process. (One can also consider every atomic sentence as representing one observation, but this is a mere technicality without much philosophical significance.)

Clearly the logical space in this semantics cannot be based on the assumption that the sequence will stop. Each domain $D(n)$ must be included in the semantics, because it is possible that the sequence will verifiably have observations of at least n individuals.

In each initial segment w_i of the observation sequence α_o, it is possible that the observations either continue or stop. There is thus a possibility of a verification at any given initial segment w_i that the sequence will continue, in the sense of a logical possibility.

Suppose w_i is verifiably an initial segment of α_o. Then a given extension w_{i+1} of w_i is a possibly verifiable initial segment of α_o, in the sense that there is a logical possibility that w_{i+1} will be verified as an initial segment of α_o. Moreover, if w_{i+1} is such a segment, then a given extension w_{i+2} of w_{i+1} is also a possibly verifiable initial segment of α_o. Hence, if w is an initial segment of α_o, then w_{i+2} is a possibly verifiable initial segment of α_o. It is thus clear that any finite extension of w is a possibly verifiable initial segment of α_o in the sense of logical possibility of verifiable truth; although it is not possible to verify the infinity of α_o, the possibility of its infinity can be verified. Moreover, even if one cannot verify some sentence S in every initial segment, its possible truth in each initial segment may be verifiable.[3]

This reveals the infinite character of the space of logical possibilities for observations clearly: there can be no upper bound in the length of the sequences which

[3]This comes close to the concept of possibility or consistency of a proposition in a choice sequence presented by Per Martin-Löf in a series of lectures at Stockholm University during 1990-1991.

can possibly be verifiable initial segments of α_o. The sequence α_o has a possibly verifiable extension at every finite stage, unless α_o has stopped for good. Hence, α_o is possibly infinite in the sense that it may never be known to have stopped.

Consider now the consecutive extensions of the initial segment w_i of α_o (where w_{i+1} is an extension of w_i etc). Clearly there is an infinite number of possible consecutive extensions which form an infinite sequence of state descriptions. But this sequence itself does not represent a possible world in the sense of verifiably corresponding to a lawless observation sequence; it is rather an infinite totality of possible or potential observations, which can never be simultaneously realized (cf. also section 2.5.5).

The section below will elaborate how the infinite sequences of possible observations figure in defining constructive truth in infinity.

5.3.3 Constructive truth and probability in the semantics of observation sequences

As stated above, the sequence α_o is possibly infinite. If S is true in α_o, it is true in a possibly infinite sequence.

If S is true in all the extensions of the initial segment w_i of α_o (which are all finite), it is verifiable in α_o – in the sense of finite truth. The idea is that no matter which finite state description α_o turns out to be, S is true in it. Verifiability in α_o thus means logical necessity in finite state descriptions under condition w_i.

However, as seen from the discussion in 4.1.3.2, this kind of concept of verifiability as necessity is not a reasonable concept of constructive truth. Moreover, it does not yet explain truth in infinity because it refers only to truth in finite extensions of w_i.

Let us assume that classical statements about the future behaviour of α_o can be made; i.e., some S can be classically true or false in α_o. What does it then mean to say that S is classically true in α_o?

Take first the case that α_o stops at some finite stage. Then S is obviously true in α_o iff it is true in that finite state description.

Next take the case that α_o in fact continues forever. In this case, S is clearly true in α_o iff it is true in that infinite sequence which is referred to by α_o. Classical truth does not seem to pose any problems here since it can be defined without being able to point out α_o in its entirety. It is evident that, however, in the case of constructive truth, one needs to refer to several finite sequences since one does not know when α_o will stop, if it will stop at all.

The logical space of n-long sequences is denoted by $W(n)$. The logical space of all finite sequences is denoted by W_{fin}, but the domain of individuals in the space of sequences of possible observations must be D_∞, consisting of an infinite

number of individual constants. The infinite sequences of possible observations form a continuum W_∞; this, however, is not the space of possible worlds, as pointed out in section 5.3.2 above.

Although the possibly infinite observation sequences clearly do not belong to the set of finite sequences W_{fin}, the notions of constructive truth and probability for observation sequences can be explained by means of sequences in W_{fin}, as will be shown next.

The observation sequences themselves will not be explicitly defined as functions of a constructive formal language; it is simply said that they are lawless sequences of observations, in the same way as lawless choice sequences are given in constructive mathematics. Every name of an observation sequence is considered as referring simply to the output of some observational setting, which is given to us in one way or another.

It would be possible to assign a probability measure to the event that α_o continues after a given finite stage. However, in order to focus on truth and probability in infinity, this possibility will not be discussed here.

Because α_o is never verifiably infinite, S's constructive truth in α_o cannot mean its truth in an infinite state description. It is also clear that S's truth in some initial segment w_i of α_o is not sufficient for its truth in the whole of α_o (consider, for example, that S is true in w_i but false in the initial segment w_{i+1} of α_o, which is an extension of w_i).

What about defining S's truth in α_o as its truth in all the consecutive initial segments of α_o beginning from some stage w_i? This concept of truth means that the totality of consecutive observations constituting α_o satisfy S, which thus in a sense remains true according to the observations in α_o. If S is true in this sense, it would be true about something that belongs to the realm of experience, namely, about consecutive stages of α_o's development.

Although truth in the above sense does apply to constructions which are to some extent observable (the finite initial segments of α_o), it is not possible in the general case to verify or falsify S about the possibly infinitely proceeding α_o. Because of this verification problem in observation sequences, truth in this sense cannot be a constructive notion. Instead, the constructive truth of S in infinity will be defined as truth in consecutive initial segments of an infinite (law-like) sequence of possible observations. Such a sequence represents the provable *possibility* that S is true at each finite stage of the development of the lawless observation sequence α_o (cf. section 5.3.2 above).[4]

[4]Note that requiring mere non-falsifiability of S instead of its truth in the consecutive initial segments of an infinite sequence of possible observations would mean that these segments are positive outcomes for the probability of both S *and* its negation (cf. p. 52). Moreover, since non-falsifiability must be defined with respect to some notion of truth in infinity, the latter remains to be defined for the infinite sequence as a whole (i.e., not by means of its initial segments), which

How is the statement "S is constructively true" then to be interpreted in the constructive semantics of observation sequences? If α_o denotes the actual world, "S is constructively true" should mean that S holds in α_o. For α_o, the expression "true" can constructively only mean verification on the basis of an initial segment. However, there may be an *a priori* restriction (cf. section 2.3) on α_o according to which admissible initial segments of α_o are only such in which S is true and which also are initial segments of sequences of possible observations in which S is true. In other words, if S holds in α_o, each initial segment of α_o will coincide only with sequences of possible observations where S is true. With this restriction, α_o is a partly law-like sequence and not an observation sequence.

The point in defining constructive empirical truth is not in *a priori* verifications. Proving something prior to experience certainly fits well in the constructive framework, but the challenge is to define constructive truth in situations where no prior to experience verification is possible.

An *a posteriori* concept of truth can be constructively defined only by reference to finite cardinalities. Hence, it can be said that S *is constructively true in* α_o's *stage* x, meaning that the x-long initial segment of α_o coincides with one of the sequences of possible observations whose consecutive initial segments satisfy S without an upper bound.

The probability of S is defined by means of the above concept of S's constructive truth in α_o's finite stage. Roughly speaking (a more precise definition will follow in the following sections), S has a probability of p in α_o iff the limit probability of α_o to coincide with one of the infinite sequences of possible observations satisfying S is p.

How can the notion of truth at a finite stage be an explication of constructive truth in infinity? The probability of S at stage x of α_o is clearly about a verifiable property of α_o, namely non-falsifiability with respect to truth in some infinite sequence of possible observations. Can the reliability of sentences be measured this way? In a sense, the answer must be positive since the property in question is the most that is knowable about the empirical world with respect to truth of any S when truth is defined by means of sequences of possible observations.

5.3.4 Asymptotic limits and truth in consecutive finite domains

This section prepares the introduction of the formal notions of extendible truth and probability. Here we discuss a particular difficulty associated with the Carnapian solution of using consecutive finite domains to define probability in infinity. This difficulty shows that Carnap's solution is not in accordance with a reasonable notion of truth in infinity without some modifications.

---

is problematic considering constructive probability in infinity, see chapter 4.

The function $r_i^*(w_n)$ yields the restriction of a n-long sequence w_n to the first i individuals, $i \leq n$ (cf. 4.1.3.1). The value of $r_i^*(w_n)$ is thus a state description of the cardinality i which is a part of a larger state description w_n of cardinality n. It is also said in this case that w_n is an *extension* of $r_i^*(w_n)$.

Assume that

$$w_n \models S, \tag{5.7}$$

and

$$(\forall w'_{n+k_i})(w_n = r_n^*(w'_{n+k_i}) \rightarrow w'_{n+k_i} \not\models S) \tag{5.8}$$

for an infinite sequence of k_i's such that $k_1 < k_2 < ....$ There is then an infinite number of cardinalities $n + k_1, n + k_2, ...$ in which S is not true for any extension of w_n. This means that w_n cannot be extended in such a way that S would be true in any series of consecutive extensions, i.e., that S is true for some w_{n+1} such that $w_n = r_n^*(w_{n+1})$, for some w_{n+2} such that $w_{n+1} = r_{n+1}^*(w_{n+2})$, etc. *ad infinitum.* In other words, if the background information consists of w_n, there is no possibility of S being true without an upper bound in consecutive state descriptions representing the flow of new information.

This kind of situation is illustrated by the following example. Suppose that the domains are nested and the individual constants in each domain are denoted by consecutive natural numbers (cf. section 5.3.1 above). Consider then a monadic language with symbols for primitive recursive functions, and the identity symbol '='. Let S' be

$$S' = (Even_D \supset A(1))\&(Odd_D \supset\sim A(1)), \tag{5.9}$$

where

$$Even_D =_{df} (\exists y)[(\forall x)(x \leq y) \supset (\exists z)(y = 2z)] \tag{5.10}$$

and

$$Odd_D =\sim Even_D. \tag{5.11}$$

The antecedent of the implication in (5.10) says that there is a largest number in the domain of individuals and the consequent that this number is even. Since the antecedent is true in every finite domain, (5.10) is true in every finite domain containing an even number of individuals. Similarly, Odd_D is true only if the consequent is false, which is the case in every finite domain containing an odd number of individuals.

In domains of even cardinalities, S' is true in those state descriptions where A holds for the individual constant 0, and in odd cardinalities in those state descriptions where $\sim A$ holds for 1, respectively.

To see this, consider first the domain $D(1)$ with the single element 1. There are two state descriptions with this domain, $A(1)$ and $\sim A(1)$. Since the cardinality of $D(1)$ is odd, the second conjunct of S' implies that $A(1)$ cannot be true if S' is to be true. Hence, S' is true in $\sim A(1)$, but not in $A(1)$.

Correspondingly, in the domain $D(2)$ with two individuals and thus even cardinality, S' is true in only those state descriptions in which $A(1)$ is true. In $D(2)$, S' is true in none of the extensions of $\sim A(1)$, and in every extension of $A(1)$. The case of $D(2)$ is similar to that of $D(1)$: S' is only true in extensions of $\sim A(1)$ but in none of $A(1)$. The case of $D(4)$ is again similar to that of $D(2)$, etc.

It can be concluded that in odd cardinalities S' is true in those state descriptions which contain $A(1)$, and in even cardinalities in those which contain $\sim A(1)$. S' thus holds in each domain in exactly $\frac{1}{2}$ of the sentences; S' thus has a limit probability of $\frac{1}{2}$.

On the other hand, it holds that

$$\begin{aligned}(\forall n)(\forall w_n)(w \vDash S' \rightarrow \\ (\forall w'_{n+1})(w = r^*_n(w') \rightarrow w' \nvDash S')),\end{aligned} \tag{5.12}$$

which means that (5.8) above holds for all n and w_n.

This example shows that the asymptotic limit probability in Carnap's sense does not correspond to a notion of truth in infinity which is based on truth in consecutive finite domains.

5.4 Extendible truth

The concept of *extendible truth* (ET) will be now introduced to solve the problem introduced above and to establish a basis for the concept of probability in infinity.

The state description w_{i+m} *extends* the truth of S from i to $i+m$, $w_{i+m} \vDash_{i,m} S$, iff

$$(\forall j)(i \leq j \leq i+m \rightarrow r^*_j(w_{i+m}) \vDash S). \tag{5.13}$$

The truth of S is *extendible* from w_i to $i+m$, $w_i \vDash^{\circ}_{i,m} S$, iff

$$(\exists w_{i+m})(w_i = r^*_i(w_{i+m}) \wedge w_{i+m} \vDash_{i,m} S). \tag{5.14}$$

The truth of S is extendible *without upper bound* from w_i, $w_i \vDash^{\circ}_{i,\forall} S$, iff

$$(\forall m)(w_i \vDash^{\circ}_{i,m} S). \tag{5.15}$$

Observe that the concept of truth in infinitely proceeding finite sequences is different from that defined in infinite sequences with $\aleph_0$ terms, although most sentences which are satisfiable in $\aleph_0$ are also satisfiable in infinitely proceeding finite sequences. One consequence of extendible truth is that sentences which are satisfiable in infinite domains only, such as

$$(\forall x) \sim R(x,x)\&(\forall x)(\forall y)(\forall z)(R(x,y)\&R(y,z) \supset R(x,z))\&(\forall y)(\exists x)R(x,y), \tag{5.16}$$

are not satisfiable at all.

Extendible truth can be given either a classical or a constructive interpretation, depending on the interpretation of $\vDash^{\circ}_{i,\forall}$ in (5.15).

The extendibility of truth without upper bound is not decidable. This entails that in the constructive interpretation it is not the case that S either is or is not extendibly true without upper bound. Classically this formulation of the principle of excluded middle can be introduced as an axiom.

5.4.1 Commutation of extendible truth with logical operations

The concept of extendible truth is non-standard in the sense of commutation with the logical operations. In other words, for most logical operations the extendible truth does not follow the conventional truth table, according to which $A \vee B$ is true if and only if A is true or B is true etc.

This fact has certain consequences for the theorems of extendible probability, although extendible probability does not violate (at least) the finite additivity axiom of the probability calculus (see section 5.4.4.1 below).

Consider first the case of *negation* $\sim$. Does extendible truth commute with $\sim$, i.e., if $\sim S$ is extendibly true, is S extendibly false and vice versa?

It is clear that even if S is extendibly false (i.e., is not extendibly true) from w_i to $i+m$, S may still be true in w_i in the ordinary sense of truth. This means that $\sim S$ is not extendibly true from w_i to $i+m$ either. It follows that extendible truth does not commute with negation $\sim$.

The case of *disjunction.* $S\vee \sim S$ is then classically and constructively valid in the sense of extendible truth since one of the disjuncts is true in the ordinary sense of truth in any given finite state description (provided, of course, that the atomic sentences are decidable). Hence, there may be cases in which $S\vee \sim S$ is extendibly true (since it is valid), but neither S nor $\sim S$ is extendibly true. It follows that extendible truth does not commute with disjunction classically or constructively.

Conjunction. If both A and B are extendibly true, $A\&B$ is extendibly true and vice versa. Hence, extendible truth commutes with conjunction.

Implication. If B is always extendibly true when A is extendibly true, i.e., the extendible truth of A entails the extendible truth of B, it may still be the case that A is not extendibly true while being true in the ordinary sense and B is not true even in the ordinary sense. If this is the case, $A \supset B$ is not true and thus cannot be extendibly true. This shows that extendible truth does not commute with implication.

Existential quantification. Consider the sentence $(\exists x)(\forall y)C(x,y)$. If it is extendibly true from w_i, $(\forall y)C(a,y)$ is true for some a in w_i. However, in the extension w_{i+1} of w_i which satisfies $(\exists x)(\forall y)C(x,y)$, it may be the case that $C(a,b)$ is false for the new constant $b \in D(i+1), b \notin D(i)$, while $C(c,y)$ is true for some $c \in D(i+1)$ and all $y \in D(i+1)$. Hence, $(\forall y)C(a,y)$ is false in w_{i+1} and is thus not extendibly true from w_i. Hence, a sentence of the form $(\exists x)P(x)$ can be extendibly true even if $P(a)$ is not extendibly true for any a, which means that extendible truth does not commute with existential quantification.

Finally, the case of *universal quantification.* If $(\forall x)P(x)$ is extendibly true from w_i, $P(a)$ must be extendibly true from w_i for all $a \in D(i)$. On the other hand, if $P(a)$ is extendibly true from w_i for all $a \in D(i)$, $P(b)$ may be false for some $b \in D(j), j > i$. Hence, $(\forall x)P(x)$ is not necessarily extendibly true from w_i. This means that extendible truth does not commute with universal quantification.

The above shows that extendible truth commutes only with conjunction.

A detailed discussion about the significance of these findings must be postponed to a further study. However, it is already clear that the concept of extendible truth challenges the traditional views about truth in infinity. Since traditional is not a synonym for intuitive, the above results do not show that extendible truth is unintuitive. Consider, for example, $A \vee B$. Perhaps the traditional requirement that if $A \vee B$ is true, then A is true or B is true could be unintuitive in a semantics of observation sequences. This question is briefly elaborated in section 5.4.4.1 below.

5.4.2 Extendible truth and truth in infinity

In what follows, a definition of sets of sequences satisfying S in the sense of extendible truth at a given cardinality will be constructed. The notion of truth of S in infinity will be based on these sets.

Consider first a family of sets of the form

$$W_u(i,S) = \{w_i \in W(i) | w_i \vDash^{\circ}_{i,\forall} S\} \tag{5.17}$$

where S is a sentence and $i \in N$. An infinite sequence satisfying S at every step can be initiated from each of the elements of $W_u(i,S)$. For an arbitrary i, $W_u(i,S)$

is the set of this kind of state description. Hence, $W_u(i, S)$ as a family of sets is not a state description or a set of them, but rather a function which provides a set of a certain kind of state description at each i.

$W_u(i + x, S)$ may always contain new state descriptions which are not extensions of elements of W_u in the preceding cardinality. The probability $P(W_u(i, S))$ is about an observation sequence α_o coinciding with such a state description in $W(i)$ from which S's truth is extendible without upper bound.

Correspondingly, $P(W_u(i + 1, S))$ measures this probability for $W(i + 1)$. The limit of $P(W_u(i, S))$ (when $i \to \infty$) can be positive even if the probability of the extensions of all $w_i \in W_u(i, S)$ satisfying S tends to zero. In other words, the limit of $P(W_u(i, S))$ is not about S being constantly true without upper bound starting from some cardinality; it is about S being true so that its truth *can* be extended without an upper bound. Hence, the limit of $P(W_u(i, S))$ is that of the probability of S's *possible* truth in infinity.

It would thus not be correct to use the limit of $P(W_u(i, S))$ to represent the probability of extendible truth of S in α_o. Instead, the following approach is preferred.

Let us define a set of the form[5]

$$\emptyset \neq F_R(i, w_i, S, x) \subseteq W(i + x), \tag{5.18}$$

which for a given state description w_i and cardinality $i + x$ consists of the extensions of $w_i = r_i^*(w_{i+x})$ in W_{i+x} for which it holds that every $w_{i+x} \in F_R(i, w_i, S, x)$ extends S's truth from i,

$$(\forall w_{i+x} \in F_R(i, w_i, S, x))(w_{i+x} \vDash_{i,x} S). \tag{5.19}$$

However, this definition does not yet guarantee that the truth of S is *extendible from* every member of F_R. The extendibility property holds by the above definitions only from w_i up to every member of F_R, which means that F_R does not necessarily contain only state descriptions from which S's truth is extendible without upper bound.

Hence, one must add a further recursive requirement for F_R:

$$\begin{aligned}(\forall w_{i+x} \in F_R(i, w_i, S, x))(\exists w_{i+x+1} \in F_R(i, w_i, S, x + 1)) \\ (r^*_{i+x}(w_{i+x+1}) = w_{i+x}).\end{aligned} \tag{5.20}$$

A state description can thus only belong to F_R if S's truth is extendible from it without an upper bound.

The constructive interpretation of F_R requires that one can actually point out the extensions of w_{i+x}, i.e., that they exist provably. In other words, a given w_{i+x}

[5] F and R standing for "finite" and "range" respectively.

cannot belong to $F_R(i, w_i, S, x)$ unless it is provable that it has extensions without upper limit which satisfy S. This means that it is not in general decidable whether a given w_{i+x} belongs to $F_R(i, w_i, S, x)$ or not.

Since this is not decidable, it could be considered that F_R is not eligible to be a set in constructive mathematics (for example, in Martin-Löf's type theory, it is required that $\in$ be an effectively computable relation). In this case, one could refer to the collection of elements which satisfy the defining conditions of F_R with a particular expression but refrain from calling it a set. In any case, the decidable notion of set is very restrictive considering the formalization of ranges of sentences in infinite domains (or infinite series of consecutive finite domains). It is therefore suggested that a constructive formalization of state description semantics should allow for undecidable sets of state descriptions. In spite of this, the positive claim that a sentence S is true in w (i.e., w belongs to the range of S) is only constructively justified if it is verifiable.

It can be said that the infinitely proceeding sequences generated by consecutive applications of F_R satisfy S (make S true) in infinity in the sense of sequences of possible observations; the function F_R describes the ways an observation sequence satisfying S might proceed from some initial state description. The elements of $F_R(i, w_i, S, x)$ for a given x are thus initial segments of sequences of possible observations.

Note that the constructive interpretation of F_R requires that the principle of mathematical induction should be applicable to infinite sequences whose terms are interpreted as possible observations (cf. section 2.5.2 above).

5.4.3 Extendible truth as a foundation of probability

It was shown in section 5.2 that in predicate logic without identity or functions there is an effective method for extending the truth of a sentence from a given state description to any larger one. This means that if $w_i \vDash S$, the truth of S can be effectively extended without upper bound. Hence, in predicate logic without identity or functions in the object language, classical and constructive extendible truth in the sense of (5.15) coincide.

To illustrate the situation in predicate logic without identity, consider the following. One can effectively find the elements of $W(i + x)$ which extend S's truth from some given $w_i \in W(i)$. If S is true in some w_{i+x}, its truth is clearly extendible to $x + i + 1$. Hence, one can effectively find the elements of any $F_R(i, w_i, S, x)$.

The underlying logical space associated with the concept of extendible truth must be the space of infinite sequences of finite state descriptions. This can be represented by the Cantor space or any suitable tree structure with nodes and branching sequences as in the infinite state descriptions. The nodes of the sequences of

the tree are now considered to be finite state descriptions instead of initial segments of some infinite state descriptions. The constructive interpretation of the tree is by means of lawless sequences (cf. section 2.3). However, the concepts of extendible truth and probability as defined above do not need an explicit definition of an infinite logical space of infinite sequences of finite state descriptions; as in Carnap 1962, one can manage with nested finite logical spaces of finite state descriptions.

As explained above, the elements of $F_R(i, w_i, S, x)$ represent initial segments of infinite sequences of possible observations. Because any finite initial segment (i.e., finite state description) can be denoted by a linguistic expression, there cannot be more than a denumerable number of initial segments. Hence, a denumerable number of infinite sequences is sufficient for producing a non-negligible asymptotic probability of finite initial segments.

Constructively, the elements of $F_R(i, w_i, S, x)$ are limited by the possibility of extending S's truth with law-like sequences in the subsequent cardinalities. To see this, assume that $w_{i+x} \in F_R(i, w_i, S, x)$ is justified because $w_{i+x} = r^*_{i+1}(\alpha)$ for a lawless α. Then there must be $y \geq 0$ such that S is constructively true in some $w_{i+x+y} \in W(i + x + y)$ and in all of its finite extensions – which are all law-like sequences.

Recall that there can be only a denumerable number of law-like sequences. There is thus only a denumerable number of infinite sequences of extensions of w_i which can be denoted. This means that most parts of the space of infinite sequences of finite state descriptions (the space of the cardinality $2^{\aleph_0}$) cannot be denoted and thus cannot figure in the constructive proof of a given sentence S. This does not limit the logical possibilities for a sentence to be true since the restriction to a denumerable number of law-like sequences does not effect the logical possibility of a given S in constructive semantics. Even if S is logically possible in the weak sense that its negation is not logically necessary, it is not possible that there is no denotable sequence making it true (cf. section 2.2 above).

5.4.4 The probability of extendible truth

The set of state descriptions in $W(i + x)$ which extend S's truth from i and from which S's truth is extendible to an arbitrarily high cardinality is

$$E_\forall(S, i, x) = \bigcup_{w_i \in W(i)} F_R(i, w_i, S, x). \tag{5.21}$$

In the set $W(i + x)$, the probability of S's extendible truth from i to an arbitrarily high cardinality is

$$P^{ET_i}(S, i, x) = P(E_\forall(S, i, x)). \tag{5.22}$$

For reasons of simplicity, it will be assumed that P corresponds to Carnap's $m^\dagger$, i.e., the equiprobability of state descriptions, but the results in what follows are likely to extend to other measures as well.

It is decidable whether a given w_{i+x} belongs to $E_\forall(S,i,x)$ if S's extendible truth without upper limit from w_{i+x} is decidable. This is the case in first-order logic without functions or identity (cf. section 5.2).

Assume a constructively defined set of real numbers, for example, by means of sequences of rational numbers, as in Bishop (1967). The limit of $P^{ET_i}(S,i,x)$ (if the limit exists), when x grows, is then the limit probability of S's extendible truth without upper bound from i:

$$P_\infty^{ET_i}(S,i) = \lim_{x\to\infty} P^{ET}(S,i,x). \tag{5.23}$$

The overall limit probability of S's extendible truth is

$$P_\infty^{ET_\infty}(S) = \lim_{i\to\infty} P_\infty^{ET_i}(S,i). \tag{5.24}$$

The limit (5.23) is zero for sentences like (5.9).

The definition of conditional extendible probability can be construed along the usual lines:

$$P_{Con}^{ET_i}(S_1,S_2,i,x) = \frac{P^{ET_i}(S_1\wedge S_2,i,x)}{P^{ET_i}(S_2,i,x)}. \tag{5.25}$$

However, a detailed treatment of conditional extendible probability will not be presented in this study.

5.4.4.1 A remark about additivity

The equation

$$P^{ET_i}(A\vee B,i,x) = P^{ET_i}(A,i,x) + P^{ET_i}(B,i,x) \quad (A\cap B = \varnothing) \tag{5.26}$$

is not a theorem. The set $E_\forall(A,i,x)$ can contain only sequences of state descriptions in which A is true; and similarly with $E_\forall(B,i,x)$. Hence, sequences of state descriptions in which A and B are alternately true are not included, although they are included in $E_\forall(A\vee B,i,x)$.

However, the extendible truth of $A\vee B$ from w_i is not the same event as the union of the events 'A is extendibly true from w_i' and 'B is extendibly true from w_i' (cf. section 5.4.1). Hence, the concept of extendible probability does not directly violate the additivity axiom

$$P(A\cup B) = P(A) + P(B) \quad (A\cap B = \varnothing). \tag{5.27}$$

One could think of extendible probability in terms of a research process. The process converges towards the truth of a sentence S iff S is confirmed by all the finite stages of the process from some stage onwards, i.e., iff S is extendibly true without upper limit from some w_i belonging to the process. Even if the research process does not converge to A or B, it can still converge to $A \vee B$: $A \vee B$ can be confirmed by all the stages of the process from w_i even if the process does not confirm either A or B at all those stages. Hence, the probability that $A \vee B$ will conform with all the data gained during the research process may be different from the joint probability of either A or B conforming with all the data.

5.4.5 The existence of the limit for P^{ET_i}

The real number L is the limit of the function $P(n)$, $\lim_{n\to\infty} P(n) = L$, if and only if for all $\epsilon > 0$ there is a natural number n_k such that

$$\text{if } n \geq n_k \text{ then } |P(n) - L| < \epsilon.$$

It is obvious that the limit for the conventional asymptotic probability P does not always exist; consider, for instance, the function $P(Even(n))$ when $n \to \infty$. Observe that since $Even(n)$ involves a binary function symbol (multiplication), this is not in contradiction with Lynch (1980), according to which asymptotic probabilities exist for first-order languages with unary function symbols when state descriptions are equally likely.

However, the extendible probability P^{ET_i} has an advantage over conventional asymptotic probability since the classical limit always exists for P^{ET_i}, as will be shown below.

To prove that a given function converges to a classical limit, one does not have to be able to generate the numerical value of the limit.

$P^{ET_i}(S, i, x)$ is a non-increasing function for each constant i and its values are in the interval $[0, 1]$. Hence, $P^{ET_i}(S, i, x)$ is bounded from below and above.

Classically, a function which is bounded from below has a greatest lower bound. A classical theorem of analysis then implies that if a is the greatest lower bound for a monotonic function like $P^{ET_i}(S, i, x)$, $P^{ET_i}(S, i, x) \to L$ for some L as $x \to \infty$.

A similar proof technique with upper bounds instead of lower bounds can be applied to the outer limit, since $P_\infty^{ET_\infty}(S)$ is clearly a non-decreasing function with an upper bound 1. This concludes the classical proof that the required classical limits always exist.

To prove the constructive existence of the limit probability, one must find it, in other words the limit must be known. The meaning of a known limit is that it is provable that a certain known L satisfies the definition of a limit. However, this

provability can be classical or constructive because one can either apply classical inference rules or restrict oneself to constructive ones. Hence, the fact that one knows that L is the classical limit does not yet entail that L constructively satisfies the definition of a limit.

5.4.6 The known limit is Markov-constructive

Is there, after all, a distinction between a known classical limit and a known constructive limit? Is a known classical limit not always a constructive limit as well?

The answer is negative in the strict sense of constructive semantics, as can be demonstrated by the reasoning below. However, as will also be seen, even a known classical limit is always constructive in a more liberal constructive sense.

The Markov school of constructive mathematics uses the intuitionistic predicate calculus with an additional principle, known as Markov's Principle (cf. Kopylov & Nogin 2001; Markov 1962; Troelstra & van Dalen 1988, pp. 203-206):

$$(\forall x)(S(x) \vee \sim S(x)) \supset (\sim\sim (\exists x)S(x) \supset (\exists x)S(x)). \tag{5.28}$$

Markov's Principle can be interpreted as follows: provided that S is decidable for a given x, it suffices for proving that an algorithm for finding an x such that $S(x)$ holds halts after a finite time to prove that the algorithm cannot possibly run forever. In other words, the existential quantifier 'a' in 'a finite time' is interpreted classically, meaning that a precise time at which the algorithm will provably terminate is not known before it actually does so.

According to Troelstra & van Dalen (1988, p. 204), the acceptability of Markov's Principle within the context of general constructivism is not all that easy to determine.

Consider now the definition of $P_{\infty}^{ET_{\infty}}(S)$. It follows from the non-decreasing property of $P_{\infty}^{ET_i}(S,i)$ that it must approach the limit $P_{\infty}^{ET_{\infty}}(S)$ from below. If this limit is interpreted classically and is known, it is known that for every ϵ an i' exists (classically!)[6] such that for every $i \geq i'$, the value of the function $P_{\infty}^{ET_i}(S,i)$ is within ϵ from $P_{\infty}^{ET_{\infty}}(S)$. One can find such an i' in a finite time by simply letting i grow because in a finite time the value of $P_{\infty}^{ET_i}(S,i)$ must be large enough (even though the value of i' cannot be specified before it is found). In other words, the search procedure for finding out the value of i' cannot possibly run forever. According to Markov's Principle, this is enough to prove that the procedure halts after a finite time. Hence, the outer limit $P_{\infty}^{ET_{\infty}}(S)$ fulfils the constructive definition of a limit in the sense of Markov's Principle.

[6]The classical existence of this i' can be proved, not only by producing the i' itself, but also by showing that the assumption that there is no such i' leads to a contradiction.

A similar argument applies to the inner limits $P_{\infty}^{ET_i}(S,i)$ since $P^{ET_i}(S,i,x)$ is a non-increasing function of x for every i. Hence, the nested limit operation $P_{\infty}^{ET\infty}(S)$ is, when known, constructive in Markov's sense.

5.4.7 Limit and verification

The aim of this section is to prove that a positive limit probability for a sentence S does not mean that S should be verifiable on the basis of some initial segment of a lawless sequence. At the outset this is not obvious. We see that extendible probability does not express the probability of the sentence being verified on the basis of finite information, but it does not follow that there actually is a non-verifiable sentence having a positive extendible probability. It is important that positive extendible probability is not limited to finitely verifiable sentences considering, for instance, the possibility of assigning positive probabilities to scientific hypotheses. Recall that one of the motivations for the notion of probability is those sentences which cannot be verified on the basis of finite evidence.

It will first be shown that the conventional Carnapian asymptotic limit probability for S may be positive even if S is not verifiable.

Consider the sentence (4.5), i.e., $(\exists x)(\forall y)C(x,y)$ again. Since the asymptotic probability of a first-order sentence without constant or function symbols is either 0 or 1 (see, e.g., Fagin 1976), the asymptotic probability of $(\exists x)(\forall y)C(x,y)$ or that of its negation must be 1. In other words, either $(\exists x)(\forall y)C(x,y)$ or its negation must have a probability of 1, although neither of them can be verified on the basis of any initial segment.

According to Liogon'kii's result (1969), the probability value of $(\exists x)(\forall y)C(x,y)$ is decidable. In order to verify constructively that one of the sentences above has a probability of 1, one should show which one. Liogon'kii's result would be handy here.

However, without assuming that Fagin's or Liogon'kii's findings hold constructively, it is impossible to assume that finitely non-verifiable sentences could not have positive extendible probabilities. The proof of this fact below will be constructive.

If the negation of Fagin's result, i.e., that the $0-1$ law does not hold for the sentences in question, was constructively provable, it would also be classically provable. But this is in contradiction with Fagin's result itself. Hence, the negation of Fagin's result is impossible.

Suppose that neither $(\exists x)(\forall y)C(x,y)$ nor its negation has an asymptotic probability of 1. Then the $0-1$ law does not hold for it. This contradicts Fagin's result, which is impossible. Hence, it cannot be assumed that neither $(\exists x)(\forall y)C(x,y)$ nor its negation has an asymptotic probability of 1.

Let us denote $(\exists x)(\forall y)C(x,y)$ by S. The argument above entails that it holds constructively that

$$\sim\sim (P_\infty(S) = 1 \vee P_\infty(\sim S) = 1) \tag{5.29}$$

where $P_\infty(S)$ denotes the conventional asymptotic probability of S.

To modify the proof about the possibility of a positive probability without verifiability in initial segments to cover extendible probability as well, further steps need to be taken. For simplicity, I will formulate the argument under the assumption that all state descriptions are equally probable (i.e., as in Carnap's $m^\dagger$); the general case would require somewhat more argumentation.

It needs to be proved that it is contradictory to assume that neither $P_\infty^{ET_\infty}(S)$ nor $P_\infty^{ET_\infty}(\sim S)$ is positive, where both S and $\sim S$ are non-verifiable.

One can prove this for an arbitrary sentence S.

It will be first proved that

$$P_\infty(S) = 1 \supset P_\infty^{ET} = 1 \tag{5.30}$$

where '$\supset$' is logical implication.

If $P_\infty(S) = 1$, for any $b < 1$ there must be i such that the probability of S in all $j \geq i$ is greater than b.

Assume first that $P_\infty(S) = 1$; then the above condition must be fulfilled. Let us choose b such that $b > \frac{1}{2}$. Let us then assume that $P_\infty^{ET_\infty}(S)$ is zero. This means that for some $j \geq i$ and for all $j' \geq j$, S is false in more than half of the state descriptions in $W(j')$. Contradiction. Hence, $P_\infty^{ET_\infty}(S)$ cannot be zero. Formally:

$$P_\infty(S) = 1 \supset P_\infty^{ET_\infty}(S) \neq 0. \tag{5.31}$$

The next step is to prove that

$$\begin{aligned} &\sim [P_\infty^{ET_\infty}(S) \neq 0 \vee P_\infty^{ET_\infty}(\sim S) \neq 0] \supset \\ &\sim (P_\infty^{ET_\infty}(S) \neq 0) \& \sim (P_\infty^{ET_\infty}(\sim S) \neq 0). \end{aligned} \tag{5.32}$$

This is easy, since (5.32) is an instance of the general theorem

$$\sim (A \vee B) \supset (\sim A \& \sim B), \tag{5.33}$$

which is constructively provable.

Let us now prove

$$\begin{aligned} \sim\sim [P_\infty(S) = 1 \vee P_\infty(\sim S) = 1] \supset \sim [&\sim (P_\infty(S) = 1) \& \\ &\sim (P_\infty(\sim S) = 1)] \end{aligned} \tag{5.34}$$

This is an instance of

$$\sim\sim (A \vee B) \supset \sim (\sim A \& \sim B), \tag{5.35}$$

which can be proved as follows. Assume

$$\sim A \& \sim B. \tag{5.36}$$

It is clear that one can derive a contradiction with (5.36) by assuming that either of the disjuncts of $A \vee B$ is true; hence, $\sim (A \vee B)$ must be true. Since assuming $\sim\sim (A \vee B)$ yields a contradiction with this, $\sim (\sim A \& \sim B)$ must be true. This proves (5.35). With (5.29), (5.34) thus entails

$$\sim [\sim (P_\infty(S) = 1) \& \sim (P_\infty(\sim S) = 1)]. \tag{5.37}$$

Suppose now that $P_\infty(S) = 1$. It follows with (5.31) that

$$P_\infty^{ET\infty}(S) \neq 0. \tag{5.38}$$

Assume also that

$$\sim [P_\infty^{ET\infty}(S) \neq 0 \vee P_\infty^{ET\infty}(\sim S) \neq 0]. \tag{5.39}$$

With (5.32), this yields

$$\sim (P_\infty^{ET\infty}(S) \neq 0) \& \sim (P_\infty^{ET\infty}(\sim S) \neq 0) \tag{5.40}$$

and further $\sim (P_\infty^{ET\infty}(S) \neq 0)$, which contradicts (5.38). Hence, $\sim (P_\infty(S) = 1)$. A similar proof with the assumption $P_\infty(\sim S) = 1$ gives $\sim (P_\infty(\sim S) = 1)$ and thus $\sim (P_\infty(S) = 1) \& \sim (P_\infty(\sim S) = 1)$, which is in contradiction with (5.37). Hence, one must deny (5.39) and thus $\sim\sim [P_\infty^{ET\infty}(S) \neq 0 \vee P_\infty^{ET\infty}(\sim S) \neq 0]$ holds. QED.

Hence denying the claim that the extendible probability of S or the extendible probability of $\sim S$ is positive leads to a contradiction. It is thus not possible to maintain that all non-verifiable sentences must have an extendible probability value of zero.

5.4.8 Calculating extendible probability

This section explores the possibility of calculating extendible probabilities. The difficulty of obtaining probabilities under the constructive interpretation of truth will be of special concern in 5.4.8.3.

5.4.8.1 Computing limits of classical extendible probability

We will now investigate how one could obtain (classical and constructive) values for extendible probabilities without an upper bound. Initially, this seems somewhat difficult.

In the case of predicate calculus without identity or function symbols, the truth of a sentence can always be constructively extended without upper bound (cf. section 5.2). It follows that if S is true in an arbitrary w_i, S's truth can be constructively extended from w_i. The probability of S's constructive extendible truth without upper bound from i is thus decidable in each $i + x$.

However, for full first-order logic, the question of the constructive or classical decidability of the set $E_\forall(S, i, x)$ in (5.21) has not been established because the prerequisite for this decidability property is the decidability of S's extendible truth without an upper bound for a given w_{i+x}.

This leads one to inquire whether there might be an easier way to compute the asymptotic limit of S's extendible probability than finding out the extendibility of S's truth without upper limit from each w_{i+x}.

The set

$$Ext(S, i, x) = \{w_{i+x} \in W(i + x) | w_{i+x} \vDash_{i,x} S\} \tag{5.41}$$

clearly incorporates at least those state descriptions in $W(i + x)$ from which S's truth can be extended without upper limit, but it may contain other state descriptions as well. Hence, the probability measure of the set $Ext(S, i, x)$ is at least $P_\infty^{ET_i}(S, i)$ in both classical and constructive interpretations of extendible probability.

It will be shown below that the classical limit probability of $Ext(S, i, y)$ when $y \to \infty$ equals the classical limit $P_\infty^{ET_i}(S, i)$ in (5.23).

Consider those $w_{i+x} \in W(i + x)$ for which $w_{i+x} \notin E_\forall(S, i, x)$. (If $w_{i+x} \in E_\forall(S, i, x)$, then w_{i+x} cannot make $Ext(S, i, x)$ and $E_\forall(S, i, x)$ differ from each other.) Then either

A) w_{i+x} extends S's truth from i but S's truth cannot be extended without upper bound from w_{i+x} or

B) w_{i+x} does not extend S's truth from i.

If B) is the case, $w \notin Ext(S, i, x)$ either. These state descriptions thus cannot make the probabilities of $E_\forall(S, i, x)$ and $Ext(S, i, x)$ differ from each other and can be left out of further deliberations.

Assume then that A) is the case for w_{i+x}. In this case w_{i+x} belongs to $Ext(S, i, x)$, which entails that $P(Ext(S, i, x)) \geq P(E_\forall(S, i, x))$. A) also entails that $w_{i+x'} \vDash\sim S$ holds for some $x' \geq x$ and for all $w_{i+x'}$ fulfilling the condition $w_{i+x} = r_{i+x}(w_{i+x'})$.

It follows from the above that in some cardinality $x'' \geq x'$ the set $Ext(S, i, x'')$ contains only extensions of state descriptions $w_{i+x} \in Ext(S, i, x)$ for which $w_{i+x} \in E_{\forall}(S, i, x)$ as well. Moreover, consider a given $w'_{i+x} \in E_{\forall}(S, i, x)$. It is possible that some extensions of w'_{i+x} in $W(i + z)$ $(z > x)$ do not satisfy S and thus do not belong to $Ext(S, i, z)$. These facts mean that $P(Ext(S, i, x''))$ cannot be greater than $P(E_{\forall}(S, i, x))$. In fact, $P(Ext(S, i, z))$ cannot be greater than $P(E_{\forall}(S, i, x))$ for any $z \geq x''$.

Hence, it holds that

$$(\exists x'')(\forall z \geq x'')(P(Ext(S, i, z)) \leq P(E_{\forall}(S, i, x)) = P^{ET_i}(S, i, x)). \quad (5.42)$$

On the other hand, because $P^{ET_i}(S, i, y)$ contains only state descriptions which satisfy S in infinity, $P(Ext(S, i, y)) \geq P^{ET_i}(S, i, y)$ for all y. It thus holds that

$$(\exists x'')(\forall z \geq x'')(P^{ET_i}(S, i, x) \geq P(Ext(S, i, z)) \geq P^{ET_i}(S, i, z)). \quad (5.43)$$

Since obviously, even when $z \geq x''$,

$$\lim_{x \to \infty} P^{ET_i}(S, i, x) = \lim_{z \to \infty} P^{ET_i}(S, i, z) = P_{\infty}^{ET_i}(S, i), \quad (5.44)$$

it must hold that

$$\lim_{x \to \infty} P(Ext(S, i, x)) = P_{\infty}^{ET_i}(S, i). \quad (5.45)$$

Equation (5.45) means that it is enough to determine the left-hand side for calculating extendible probabilities. This is a considerable advantage since $Ext(S, i, x)$ is decidable whereas $E_{\forall}(S, i, x)$ is not.

5.4.8.2 Computing limits of constructive extendible probability

In this section the finding of the preceding section will be discussed for the case of a constructive notion of asymptotic limit.

The question of the constructive validity of the argument in the previous section can be split into two questions. First, there is the question of whether the constructive limit can be calculated in the way suggested. Second, the constructive validity of the reasoning itself is to be examined.

Let us analyse the above proof step by step.

The known value of $E_{\forall}(S, i, x)$ can only differ from the known value of $Ext(S, i, x)$ when it is known that A) holds for certain w_{i+x} (see section 5.4.8.1 above). But A) holds knowably only if it is known that S's truth is not extendible from w_{i+x} without upper bound.

Applying the reasoning of the previous section (but without assuming knowledge about a particular x'') we get

$$\sim\sim (\exists x'')(\forall z \geq x'')(P^{ET_i}(S,i,x) \geq P(Ext(S,i,z)) \geq P^{ET_i}(S,i,z)). \quad (5.46)$$

This holds constructively even if no x'' can be effectively found. It is then clear that

$$\lim_{x\to\infty} P(Ext(S,i,x)) \neq P_{\infty}^{ET_i}(S,i) \quad (5.47)$$

cannot be true and thus (5.45) holds. This shows that (5.45) holds constructively when the limits exist constructively.

However, the motivation for deriving (5.45) above was to be able to use the value $Ext(S,i,x)$ instead of the undecidable $E_{\forall}(S,i,x)$. Precisely the situations in which it is *not* known whether some $w_{i+x} \in Ext(S,i,x)$ also belongs to $E_{\forall}(S,i,x)$ are the reason why (5.45) was attempted. This means that situations in which either $w_{i+x} \in E_{\forall}(S,i,x)$ or $w_{i+x} \notin E_{\forall}(S,i,x)$ is known are not the crux of the matter, the whole idea being able to avoid the need to know whether $w_{i+x} \in E_{\forall}(S,i,x)$ holds when calculating S's extendible probability. In the constructive version of the proof, one thus cannot rely on assuming that $w_{i+x} \in E_{\forall}(S,i,x) \vee w_{i+x} \notin E_{\forall}(S,i,x)$ holds constructively.

Result (5.45) shows that even if one found out for every w_{i+x} whether $w_{i+x} \in P(E_{\forall}(S,i,x))$ or not, it could not change the limit probability of $E_{\forall}(S,i,x))$, provided that the latter exists. If the limit exists constructively (which requires that it be known), it must be the same as the limit of $P(Ext(S,i,x))$.

On the other hand, (5.45) is not, after all, constructively valid. Computing the left side of the equation does not mean that the value of the right side is constructively established. Establishing the value of the right side is not trivial since the constructive extendible truth of S from w_{i+x} requires a function that effectively extends S's truth without upper bound, and it is not an effectively solvable task to find such a function in full first-order logic. For first-order logic without functions or an identity symbol, it was shown in 5.2 that in for every cardinality i in which a certain w_i satisfies S, this kind of function can be constructed. Hence, for this segment of first-order logic, (5.45) holds constructively without qualification.

In what follows it will be proved that if the value of $\lim_{x\to\infty} P(Ext(S,i,x))$ exists constructively, it is possible to approximate the constructive value of every $P_{\infty}^{ET_i}(S,i)$ with arbitrary precision.

5.4.8.3 Producing the initial segments

Calculating extendible probabilities in the constructive interpretation of extendible truth is especially difficult because discovering which state descriptions meet the requirement (5.20) of $F_R(i,w_i,S,x)$ is a non-trivial matter.

It will now be argued that, when

$$\lim_{x\to\infty} P(Ext(S,i,x)) > 0 \tag{5.48}$$

holds constructively, there is a method for producing almost all the initial segments satisfying the definition of extendible truth without upper bound and contributing to the limit probability of S in the sense that the joint probability of the extensions does not tend to zero. For L to represent constructive extendible probability, it is required at each finite cardinality that one can point out the state descriptions contributing to the limit in the sense that S's truth is extendible without an upper bound, which is the condition (5.22) above, whereas finding state descriptions whose extensions do not have a positive limit probability is not essential for considering L as the limit of constructive extendible probability of S.

The argument below focuses first on the cardinality i.

Suppose the constructive limit of $P(Ext(S,i,x))$ is $L \neq 0$. Let the set of state descriptions in $W(i)$ from which S's truth is extendible to $i+x$ be

$$Res(S,i,x) = \{w_i \in W(i) | w_i \vDash^{\circ}_{i,x} S\} \tag{5.49}$$

and let the set of state descriptions in $W(i+x)$ which extend S's truth from w_i be

$$\begin{gathered} E_I(S,i,x,w_i) = \\ \{w_{i+x} | w_{i+x} \in W(i+x) \wedge w_{i+x} \vDash_{i,x} S \wedge w_i = r_i^*(w_{i+x})\}. \end{gathered} \tag{5.50}$$

It is clear that

$$Ext(S,i,x) = \sum_{w_i} E_I(S,i,x,w_i). \tag{5.51}$$

Then

$$\lim_{x\to\infty} \sum_{w_i} P(E_I(S,i,x,w_i)) = \lim_{x\to\infty} P(Ext(S,i,x)) = L. \tag{5.52}$$

The fraction

$$\frac{L}{|Res(S,i,x)|} \tag{5.53}$$

in which $|Res(S,i,x)|$ is the number of state descriptions in $Res(S,i,x)$, is the average limit probability of extensions of elements in $Res(S,i,x)$. In other words, on average, the extensions of each w_i in $Res(S,i,x)$ have a combined limit probability according to (5.53), although not all w_i in $Res(S,i,x)$ necessarily have extensions without upper limit.

$Res(S, i, x)$ is a non-increasing function of x, which means that (5.53) is non-decreasing. It is clear that the distance of $Ext(S, i, x)$ from the limit,

$$D(S, i, x, L) = P(Ext(S, i, x)) - L \tag{5.54}$$

may be arbitrarily small when x increases (although not necessarily zero). Then x' can be effectively found (recall that L is a constructive limit) such that

$$\frac{L}{|Res(S, i, x')|} > D(S, i, x', L). \tag{5.55}$$

Recall that $Res(S, i, x')$ consists of state descriptions from which S's truth is extendible to $i + x'$. The combined probability of extensions of every $w_i \in Res(S, i, x')$ must have a positive probability at x'. If $D(S, i, x', L) = 0$, this probability cannot decrease, which means that the desired result has been established, i.e., we have pointed out those w_i whose extensions satisfying S have a positive limit.

In the general case, $P(Ext(S, i, x))$ approaches L asymptotically. If (5.55) holds, then the probability of extensions in $W(i + x')$ of the average element of $Res(S, i, x')$ is greater than $D(S, i, x', L)$. The magnitude of $D(S, i, x', L)$ is thus comparable to that of the average element in $Res(S, i, x')$. Since there are only a finite number of elements in $Res(S, i, x')$, those for which it holds that

$$P(E_I(S, i, x', w_i)) > D(S, i, x', L) \tag{5.56}$$

can be effectively pointed out. If (5.56) holds for w_i,

$$P(E_I(S, i, x', w_i)) - D(S, i, x', L) \tag{5.57}$$

is a lower limit for the value of $P(E_I(S, i, x, w_i)$ when $x \to \infty$.

This method of finding the relevant elements of W_i can be made more complete in the manner which will be outlined below.

Consider those $w_i \in Res(S, i, x')$ for which (5.56) does not hold. In general, for an arbitrary x, this set is non-empty and can even consist of all $w_i \in Res(S, i, x)$. Let us denote this set for x' by

$$T(S, i, x', L). \tag{5.58}$$

If

$$P(\sum_{w_i \in T} E_I(S, i, x', w_i)) > D(S, i, x', L) \tag{5.59}$$

when $T = T(S, i, x', L)$, it is clear that the probability of extensions of all $w_i \in T(S, i, x', L)$ cannot tend to zero. The minimum of $P(T(S, i, x, L))$ when $x \to 0$ is then

$$P(\sum_{w_i \in T} E_I(S, i, x', w_i)) - D(S, i, x', L). \tag{5.60}$$

Let us consider the fraction

$$\frac{P(\sum_{w_i \in T} E_I(S, i, x', w_i)) - D(S, i, x', L)}{|T(S, i, x', L|}. \tag{5.61}$$

Since $D(S, i, x, L)$ is a decreasing function of x, one can effectively find an x'' for which

$$D(S, i, x'', L) < \frac{P(\sum_{w_i \in T} E_I(S, i, x', w_i)) - D(S, i, x', L)}{|T(S, i, x', L|}. \tag{5.62}$$

Hence, for the average $w_i \in T(S, i, x', L)$,

$$P(E_I(S, i, x'', w_i)) > D(S, i, x'', L), \tag{5.63}$$

i.e., (5.56) holds when x' is substituted by x''. One can thus effectively find at least one $w_i \in T(S, i, x', L)$ for which (5.63) holds.

This reasoning can be repeated for $T(S, i, x'', L)$ effectively until (5.59) does not hold for some $y > x'$. The value $D(S, i, y, L)$ is the proportion of the limit value L for which it cannot necessarily be pointed out which $w_i \in Res(S, i, x)$ contribute to this part of the limit. In other words, $D(S, i, y, L)$ does not necessarily manifest constructive extendible probability, i.e., probability in the sense of constructive extendible truth. However, since $D(S, i, y, L)$ is a decreasing function of y where y can be effectively selected to yield a given value for $D(S, i, y, L)$, it is possible to reduce the non-constructive proportion of the limit probability down to an arbitrarily small number.

In general, similar reasoning applies for the set $Res(S, i + 1, x)$.

The above, however, not enough to show that S's truth is constructively extendible from a particular $w_i \in Res(S, i, x)$ (i arbitrary) without an upper bound. It has only been shown that the assumption that S's truth is not extendible from a subset of $Res(S, i, x)$ results in a contradiction. For constructive extendible truth, one must also suggest a method for finding at least one extension for the relevant $w_i \in Res(S, i, x)$.

Suppose that

$$P(E_I(S, i, u, w_i)) > D(S, i, u, L) \tag{5.64}$$

holds for some w_i and $u > i$. Consider now the set of extensions of w_i in W_{i+1}, say w_{i+1}^1 and w_{i+1}^2. It clearly holds that

$$P(E_I(S, i, u, w_i)) = \tag{5.65}$$
$$P(E_I(S, i+1, u, w_{i+1}^1)) + P(E_I(S, i+1, u, w_{i+1}^2)). \tag{5.66}$$

If either $P(E_I(S, i+1, u, w_{i+1}^1)) > D(S, i, u, L)$ or $P(E_I(S, i+1, u, w_{i+1}^2)) > D(S, i, u, L)$ holds, then w_i has an extension in $W(i+1)$ satisfying S, QED. On the other hand, if neither of these inequalities hold, consider the least possible limit value of $P(E_I(S, i, x, w_i))$ when $x \to \infty$,

$$P(E_I(S, i, u, w_i)) - D(S, i, u, L). \tag{5.67}$$

For some z which can be effectively found it holds that

$$P(E_I(S, i, u, w_i)) - D(S, i, u, L) > \tag{5.68}$$
$$D(S, i, z, L). \tag{5.69}$$

If now either $P(E_I(S, i+1, z, w_{i+1}^1)) > D(S, i, z, L)$ or $P(E_I(S, i+1, z, w_{i+1}^2)) > D(S, i, z, L)$ holds, then w_i has the required extension, QED. On the other hand, if $P(E_I(S, i+1, z, w_{i+1}^1)) \leq D(S, i, z, L)$ and $P(E_I(S, i+1, z, w_{i+1}^2)) \leq D(S, i, z, L)$, the limits of both $P(E_I(S, i+1, z, w_{i+1}^1))$ and $P(E_I(S, i+1, z, w_{i+1}^2))$ must be positive when $x \to \infty$ since otherwise the limit of $P(E_I(S, i, x, w_i))$ $(x \to \infty)$ would be less than (5.67), i.e., less than its minimum. QED.

Hence, there is also a method for finding out at least one extension for each i and w_i for which (5.56) holds.

It has been proved in this section that one can find almost all w_i for a given i of which it cannot be assumed that their extensions satisfying S do not have a positive limit probability. There is a method for providing at least one required extension for each such w_i. The methods introduced in this section do not provide the limit probability of the extensions of any particular w_i satisfying S in the sense of extendible truth, but it has been shown that the sum of such limit probabilities must be close to L.

Chapter 6

Second-order probabilities

This chapter starts the discussion about Carnap's inductive logic from the point of view of the problem of induction. The most difficult problem with the justification of Bayesian inference also troubles inductive logic, namely, the dependence of probability inferences on the choice of the prior distribution. The question is which prior distribution yields the most accurate predictions. One cannot exclude even prior distributions which do not take the effect of evidence into account. This is the crux of the problem of induction as formulated within the framework of general Bayesian inference.

Prior distributions are called inductive methods in Carnap's generalization of inductive logic (1952), which defines a whole continuum of inductive methods. Hence, the problem of choosing the prior distribution amounts to the problem of choosing the inductive method from this continuum.

The customary way to let evidence affect probabilities is by introducing the dyadic probability function with the conditionalization rule

$$P(h|e) = \frac{P(h\&e)}{P(e)}. \tag{6.1}$$

Conditionalization rule (6.1) in fact defines a probability function $P_e(h) = P(h|e)$ which is different from the prior probability function $P(h)$. However, since $P_e(h)$ is defined by using P, the probabilities that are obtained by applying $P_e(h)$ depend on the choice of P.

A possible solution to the problem of choosing between the available prior distributions is to assign probability values on the prior distributions themselves. This kind of second-order probability function reflects uncertainty concerning the probability values of the first order, i.e., whether the probabilities determined by applying the appropriate prior distribution are really the "true" probabilities.[1] The

[1] Carnap (1952) defines the notion of optimum inductive method for a given state description.

idea is that the choice of a second-order probability distribution can perhaps be argued for more objectively than the choice of the first-order probability distribution. Moreover, one could update the probability of the inductive method itself by evidence instead of only updating the probabilities of sentences. This would mean that Bayesian updating according to conditionalization rule (6.1) would be extended to the level of second-order probabilities, provided that the technical details of this conditionalization can be sorted out.

Apart from resorting to second-order probabilities, there is also another possible strategy for adjusting the prior probability distribution. This is to adjust the prior probability distribution with evidence in some other way than using conditionalization rule (6.1). This would amount to not just updating the probabilities of sentences by using a specified prior probability distribution $P(h)$ in (6.1), but also adjusting *which* probability distribution $P(h)$ will be used in (6.1). One usually calls this kind of updating method non-Bayesian because it deviates from the customary Bayesian rule (6.1).

This chapter first discusses some problems present in the second-order probability approach. These considerations, which so far have not appeared in this form in the literature, will lead the present author to favour a non-Bayesian way of adjusting the prior distributions (i.e., the inductive methods). Chapter 7 presents a such a non-Bayesian method for Carnap's inductive logic and an argument for using it in a situation where there is no criterion available for choosing between the different inductive methods.

6.1 Interpreting second-order probabilities

One possible way to make adjustments to inductive methods is to change their probabilities instead of fully committing to another method. Assignment of probabilities to inductive methods, however, gives rise to a question of a semantic nature: what is really meant by probability in this case? The question is important since the main objective of inductive logic is to provide a satisfactory interpretation of epistemic probability.

Probabilities of inductive methods cannot be determined in the same fashion as probabilities of sentences of the object language. This is explained as follows. Sentences can be assigned meanings, in which case they describe the world. One can assign probability values to sentences by means of applying a probability function to them, which means assigning such values to various propositional contents. Probabilities in inductive logic are considered as probabilities of obtaining of certain states of affairs, which are described by sentences. Therefore inductive

The true probability can be defined as the one given by the optimum inductive method. This will be discussed in more detail in what follows.

methods can be considered as functions from sentences to real numbers. It does not easily make sense to apply a second-order probability function to an inductive method because such a method does not describe a state of affairs in the sense of a sentence. A state of affairs can obtain – in this case the corresponding sentence is true – but it is not clear what it would mean that an inductive method obtains. Provided that second-order probabilities deal with inductive methods in one sense or another, it must be explained what states of affairs the inductive methods can be involved in such a way that probabilities can be assigned to these states of affairs.

Since parameter λ ranges over the continuum of inductive methods, its value uniquely defines an inductive method. One option is to consider second-order probabilities as degrees of belief concerning the optimum value of λ. On this view, the probability assigned to a particular value of λ, say λ', is interpreted as the epistemic probability of the state of affairs that λ' is the optimum value of λ in the state description which corresponds to the actual world.

This way of formulating the issue gives rise to a further question concerning the meaning of the word 'optimum' in this context. What are the criteria to be used in judging whether a method is the optimum one? Is there always a unique optimum method?

6.2 Subjective second-order probabilities

If it is held that one always knows one's own beliefs, subjective second order probabilities become absurd since they would always be either 0 or 1 for each first-order probability. If my degree of belief in S is p, then I know that I believe in S to the degree p so that the second-order degree of belief about my degree of belief in S being p must be one.

This view has been criticized e.g. by Skyrms (1980, p. 114) and Logue (1991, p. 158). One can, for instance, interpret degrees of belief as inclinations to behave in a certain way (in a betting situation, for example), and such inclinations are not necessarily accessible to the cognitive agent at all times since one does not necessarily know with certainty the contents of one's own mind. This argument would favour the view that subjective second-order probabilities cannot be dismissed with the triviality argument above.

However, even if second-order probabilities were feasible in the subjective interpretation of probability, second-order subjective probabilities can only be interpreted as probabilities about one's own beliefs and behaviour. There is no question about the closeness of some first-order probabilities to the true or objective probability because such a thing as true probability is supposed not to exist.

Whether second-order probabilities are justified or not in the subjective interpretation of probability is a research topic in its own right, but this study is more

directed towards finding a way of justifying a given probability distribution on the basis of some objective grounds. The motivation is thus to eliminate the subjective elements from probability assignments as much as possible, and it is difficult to see how considering second-order probabilities as subjective probabilities about subjective probabilities would contribute to such aims.

6.3 The optimum method

A more objectivistic answer to the question of interpreting second-order probabilities can be outlined by using the concept of optimum inductive method in Carnap's λ-continuum of inductive methods (1952). One can stipulate that λ_i is the optimum method if the actual (possibly unknown) state description is such that λ_i yields the most successful predictions. This section will evaluate whether a second-order probability distribution over all possible optimum values in Carnap's λ-continuum is feasible. Since each value of λ is possibly the optimum value in the (unknown) actual state description, the second-order probability distribution should be defined over the entire λ-continuum from 0 to ∞. However, a slightly different approach will be used.

Carnap (1952) derives a formula for the optimum method in a given state description which is based on numbers of individuals satisfying various *Q-predicates*.

In monadic predicate logic, a Q-predicate is a conjunction which contains either the predicate itself or its negation for each atomic predicate.[2] Hence, a Q-predicate says which atomic properties an individual possesses and which it does not. Each Q-predicate can thus be considered as a class of individuals (or individual constants) of a particular type. The set of Q-predicates incorporates all possible combinations of atomic properties an individual can have. The number of Q-predicates in the language is 2^π, where π is the number of atomic predicates of the language. Hence, individual constants can be divided into 2^π different classes. Following Carnap (1952), the number of Q-predicates will be denoted by κ in what follows.

The *degree of order* of a state description is given as follows:

$$\sum_i r_i^2 \tag{6.2}$$

where r_i is the relative frequency of the Q-predicate Q_i in the state description in question and the summation goes to κ.[3]

[2]Carnap's systems in 1962 and 1952 use monadic language only.

[3]Carnap (1952, p. 66) briefly discusses the degree of order (degree of uniformity, degree of homogeneity) of the universe. One can conclude from the text that the sum above (which is an

If the universe of discourse has N individuals and, according to the state description w, all of them satisfy the same Q-predicate, w's degree of order is maximal. On the other hand, if the individuals are distributed evenly among all possible Q-predicates, w's degree of order is minimal. (Cf. Carnap 1952, p. 66.)[4]

According to Carnap (1952), there is a unique inductive method for a given state description that yields the best estimates about the relative frequencies of the Q-predicates on the basis of any sample which is small compared to the total number of individuals N.[5]

---

essential part of the so-called Gini diversity index, see, e.g., Festa 1994) is not meant there as a definition, but merely as a quantity that reflects the degree of order. According to a footnote on that page, the concept of degree of order will be discussed in a forthcoming article. However the only published work known to the present author where the degree of order is discussed is Carnap (1977). The degree of order is there defined in a slightly different way and is referred to as the traditional concept of degree of order. In Carnap (1977), it is also mentioned that the traditional concept is examined elsewhere in more detail. Going through Carnap's unpublished manuscripts might shed more light on this issue.

[4]However, a proof of these facts is not available in Carnap (1952). Let us first prove that 1) if all individuals are concentrated on one Q-predicate, the degree of order is 1, and 2) if the individuals are evenly distributed over the Q-predicates, the degree of order is $\frac{1}{\kappa}$. Number 1) is trivial: if for some i, $r_i = 1$, then $\sum_i r_i^2 = 1$. For number 2): if for every i, $r_i = \frac{1}{\kappa}$, then clearly $\sum_i r_i^2 = \sum_i \frac{1}{\kappa^2} = \frac{1}{\kappa}$.

The final step in the proof is to show that these values are in fact the minimum and maximum for the degree of order and that they are unique in the sense that each of them is reached with precisely one statistical distribution of individuals among the Q-predicates. (The proof of this step was suggested to me by Theo Kuipers.) This requirement is implicitly assumed in Carnap (1952) when it is said that the minimum is reached when the individuals are evenly distributed and the maximum is reached when the individuals are uniformly distributed. If the minimum and maximum values of $\sum_i r_i^2$ were *also* reached with some other distributions of individuals than maximally heterogeneous and maximally homogeneous ones, it would not make much sense to regard $\sum_i r_i^2$ as expressing the degree of order of the universe.

Observe first that

$$0 \leq r_i \leq 1, \text{for all } i = 1, 2, ..., \kappa \tag{6.3}$$

and

$$\sum_i r_i = 1. \tag{6.4}$$

Observe then that $r_i^2 < r_i$ for every i if the universe is not homogeneous and $(r_i - \frac{1}{\kappa})^2 > 0$ for some i if the universe is not heterogeneous. For the maximum it then follows from (6.4) that $\sum_i r_i^2 < \sum_i r_i = 1$, QED. The minimum goes as follows:

$$0 < \sum_i (r_i - \frac{1}{\kappa})^2 = \sum_i r_i^2 - \frac{2}{\kappa} \sum_i r_i + \frac{\kappa}{\kappa^2}. \tag{6.5}$$

By (6.4), this yields $0 < \sum_i r_i^2 - \frac{1}{\kappa} \Leftrightarrow \frac{1}{\kappa} < \sum_i r_i^2$, QED.

[5]The best estimate is defined as the minimum of the mean square error of the estimate function

Following Carnap (1952, p. 69), λ^{Δ} will be used to denote the optimum inductive method. The value of λ^{Δ} depends on the variable r and the parameter π, standing for the degree of order of the state description and the number of primitive predicates in the language respectively.

λ^{Δ} can be treated as the function $\lambda^{\Delta}(r, \kappa)$. Carnap (1952) derives the following equation for $\lambda^{\Delta}(r, \kappa)$:

$$\lambda^{\Delta}(r, \kappa) = \frac{1 - r}{r - \frac{1}{\kappa}}. \tag{6.6}$$

It is easily seen from the above that the function $\lambda^{\Delta}(r, \kappa)$ is reversible. Hence, for each optimum method there is precisely one degree of order for which it is the optimum method and vice versa.[6]

Let us denote by $Desc(r)$ the set of state descriptions whose degree of order is r and let w_a denote the unknown state description corresponding to the actual state of affairs. If a particular λ' is the optimum method for w_a, the latter must manifest the degree of order associated with λ', i.e., w_a must belong to $Desc(r)$. In this case, λ' is the optimum method for inductive inference in w_a.

What does this mean constructively? If it is constructively justified to say that λ is the optimum method, this fact should be knowable. Hence, in this case it should be knowable that

$$w_a \in Desc(r). \tag{6.7}$$

In other words, whenever saying that λ is the optimum method in w_a is constructively justified, the degree of order of w_a should be knowable. The consequences of this observation will be discussed in the section below.

6.3.1 The optimum method and infinite state descriptions

This section will discuss a difficulty of defining a second-order probability distribution over the values of λ as representing the optimum method in the sense of constructive semantics. The difficulty arises when dealing with infinite state descriptions, where no decidable truth is available in general.

Suppose that the degree of order is somehow constructively defined for infinite state descriptions. In general, one not only does not know, but cannot even find out the degree of order of the state description corresponding to the actual world. This is the case in infinite state descriptions, giving rise to the problem of choosing

given by the inductive method in question, see Carnap (1952, p. 61-62). This is discussed in more detail in chapter 7.

[6] If the degree of order is $\frac{1}{\kappa}$, then formula (6.6) yields $\lambda^{\Delta} \to \infty$ when $r \to \frac{1}{\kappa}$ (cf. Carnap 1952, p. 69).

the optimum inductive method. If one knew the degree of order of the actual state description, one would also know which inductive method is the optimum one.

Consider now the constructive meaning of "λ is the optimum method" as the knowability of this fact. The probability of "λ is the optimum method" would mean the probability of getting to know that λ is the optimum method. But it is not the case that "λ is the optimum method" would always hold constructively for some value of λ since with infinite state descriptions it is not even always possible in principle to establish that a particular inductive method is the optimum one. Even if this was possible for a given infinite state description, it would not be possible to tell whether this state description corresponds to the actual world.

What does this mean for second-order probabilities over the λ-continuum? Observe first that in fact one can more easily define the probabilities over the values of the degree of order since they are in one-to-one correspondence with optimum values of λ (see above).

The degree of order r can be treated has a continuous random variable ranging between $\frac{1}{\kappa}$ and 1. In this treatment, each value of r would denote the event that the corresponding value of λ^{Δ} according to (6.6) is the optimum value for λ and these events (values of random variable r) could be assigned a density function of a continuous distribution, corresponding to the probability function over the values of a discrete random variable.

However, in constructive interpretation r cannot be this kind of random variable, the reason being that there is no density function $f(r)$ defined over the values of λ which would fulfil the condition

$$\int_{\frac{1}{\kappa}}^{1} f(r)dr = 1 \tag{6.8}$$

which is required from any density function. This is because when interpreted constructively, the expression (6.8) would mean that the degree of order has a knowable value between $\frac{1}{\kappa}$ and 1 with a probability of 1. In other words, the degree of order would always be knowable, which cannot be the case in semantics with infinite state descriptions.

6.3.2 The optimum method with finite state descriptions

Recall now the connection (6.6) between the optimum inductive method and the degree of order of the state description. Assigning a second-order probability distribution to the λ-continuum would in fact amount to assigning a probability distribution over degrees of order. It is conceivable that one could work out a constructively acceptable definition of second-order probabilities over degrees of order using the notion of extendible probability, which is based on finite state

descriptions. This would mean that the problem discussed above in section 6.3.1 could be avoided.

However, even such a treatment would not yield significant results with respect to the problem of choosing the correct inductive method. To see this, consider the following argument, which is valid in both classical and constructive semantics.

The probability of a given degree of order must equal the combined probability of the state descriptions with this degree of order. In fact, the methods in the λ-continuum define probabilities for various degrees of order by assigning probabilities to state descriptions. These are the first-order probabilities. It is not reasonable to assume that a method of assigning probabilities over degrees of order could be selected on more objective criteria at the second-order level than at the first-order level. It is thus not reasonable to expect benefits from introducing such second-order probabilities. Moreover, second-order probabilities actually become redundant if they are conceived of as probabilities over the optimality of inductive methods. This can be shown formally as follows.

Let us assume that the space of possible degrees of order is discrete. In this space not all λ-methods have a corresponding degree of order. Let $O(\lambda_i)$ denote the proposition saying that λ_i is the optimum method where i ranges over the methods corresponding to all possible degrees of order. In fact, $O(\lambda)$ can be expressed as a disjunction of state descriptions. These disjunctions are mutually exclusive (i.e., no member of any disjunction belongs to another disjunction) and incorporate all state descriptions of the language.

When $\hat{P}(O(\lambda_i))$ is the second-order probability of the optimality of λ_i and $P^i(S|O(\lambda_i))$ is the probability of S (using λ_i) under the condition that λ_i is the optimum method,

$$\hat{P}(O(\lambda_i)) \cdot P^i(S|O(\lambda_i)) \tag{6.9}$$

may be used to express the probability that S is true and λ is the optimum method. According to conditionalization rule (6.1), formula (6.9) above is equal to

$$\hat{P}(O(\lambda_i)) \cdot \frac{P^i(S\&O(\lambda_i))}{P^i(O(\lambda_i))}. \tag{6.10}$$

Observe that $O(\lambda_i)$ occurs as an argument in two probability functions, $\hat{P}$ and P^i. However, one cannot assign two different probabilities to any sentence at the same time, which means that $\hat{P}$ and P^i must denote the same function. On the other hand, it certainly does not make sense to index the second-order prior distribution over $O(\lambda_i)$ on i.

Let P replace all occurrences of probability functions in (6.11) thus yielding

$$P(O(\lambda_i)) \cdot \frac{P(S\&O(\lambda_i))}{P(O(\lambda_i))} = P(S\&O(\lambda_i)). \tag{6.11}$$

When it is not known which $O(\lambda_i)$ obtains, the overall probability of S can be calculated as the sum of all probabilities $P(S\&O(\lambda_i))$:

$$\sum_i P(S\&O(\lambda_i)) \tag{6.12}$$

Since the various $O(\lambda_i)$ exhaust all possibilities, (6.15) reduces to $P(S)$. This means that the initial problem, the problem of choosing the first-order probability distribution over the sentences of the object language, has recurred. Hence, a second-order assignment of probabilities on the optimality of inductive methods is not a solution to this problem.

6.4 The rationality of inductive methods

In this section, the second-order probabilities of inductive methods are discussed by using a more general idea of the rationality of inductive methods than their optimality in a given state description. As in the preceding section, the discussion in this section applies both in classical and constructive semantics.

It was shown above that a second-order assignment of probabilities over the optimality of inductive methods must reduce to a first-order probability assignment over the sentences of the object language. It seems, based on the observations above, that a true second-order probability distribution needs to be disconnected from the first-order distributions in the sense that the second-order distribution is not defined over the same object language sentences as the first-order one – otherwise the second-order distribution cannot differ from the first-order one. But is a true second-order probability distribution over inductive methods possible? What is the interpretation of such a distribution?

Apart from the knowable optimality of the inductive method λ, there may be other rational reasons for choosing λ. One adopts a probability distribution because it is the most rational of all the available distributions, but rationality does not necessarily mean knowable optimality. There might be some prior-to-evidence considerations indicating that some methods are more rational than others. Uncertainty about the most rational first-order distribution can be represented by the second-order probability distribution.

To avoid interference with first-order probabilities, which caused problems in section 6.3.2 above, the second-order distribution over the inductive methods should not imply that some specific probabilities must be assigned to state descriptions of the object language. The second-order distribution should merely reflect uncertainty on the issue of which one of the inductive methods is the most rational one for determining probabilities of state descriptions (and other sentences).

6.4.1 Calculating second-order conditional probabilities

This section will discuss the situation when second-order probabilities are combined with conditionalization on evidence. It will turn out that conditionalization on second-order probabilities is redundant when it comes to the conditional probabilities of hypotheses.

Let us assume, as above, that the available inductive methods can be denoted by natural numbers i. The probability of an inductive method λ_i is denoted by $\hat{P}(\lambda)$ and it is required that

$$\sum_i \hat{P}(\lambda_i) = 1. \tag{6.13}$$

If λ_i is the most rational method, the probability of S is determined according to λ_i and is denoted by $P^i(S)$.

The expressions λ_i for the inductive method and S for a sentence are not expressions of the same object language. However, the prior probability that λ_i is the most rational method and that S simultaneously obtains ("λ_i and S") can be stated as

$$\hat{P}(\lambda_i) \cdot P^i(S). \tag{6.14}$$

Here $\hat{P}$ is any prior second-order distribution defined by inductive methods and P^i a probability function corresponding to the method denoted by i.

The overall prior probability of S, $P(S)$, can be stated as follows:

$$P(S) = \sum_i \hat{P}(\lambda_i) \cdot P^i(S). \tag{6.15}$$

When second-order probabilities are involved, conditional probabilities of hypotheses can be defined with or without updating the second-order probabilities.

Consider first the case without updating second-order probabilities. Applying the conditionalization rule (6.1) on prior probabilities of the form (6.15) one could define

$$P(h|e) = \frac{P(h\&e)}{P(e)} = \frac{\sum_i \hat{P}(\lambda_i) \cdot P^i(h\&e)}{\sum_i \hat{P}(\lambda_i) \cdot P^i(e)} \tag{6.16}$$

One could also think of expressing the probability of $h\&e$ by means of using the prior probabilities of h and e. This can be done by using the following probability calculus theorem:

$$P(h\&e) = P(h) + P(e) - P(h \vee e). \tag{6.17}$$

Does the conditional probability defined by means of this rule differ from (6.16) above? This will revealed out by a direct application of (6.17). Let us check the situation by stipulating that

$$\begin{aligned} P(h\&e) &= \sum_i \hat{P}(\lambda_i) \cdot P^i(h) + \sum_i \hat{P}(\lambda_i) \cdot P^i(e) - \\ &\qquad \sum_i \hat{P}(\lambda_i) \cdot P^i(h \vee e) \\ &= \sum_i \hat{P}(\lambda_i) \cdot [P^i(h) + P^i(e) - P^i(h \vee e)] \\ &= \sum_i \hat{P}(\lambda_i) P^i(h\&e). \end{aligned} \tag{6.18}$$

It is immediately seen that this equals the numerator of (6.16), which entails that substituting $h\&e$ directly in (6.15) leads to the same result as using the formula $P(h) + P(e) - P(h \vee e)$ for $P(h\&e)$. This also demonstrates that (6.17) holds in the present formalism.

Since the inductive methods have prior second-order probabilities, the question arises whether they also have posterior second-order probabilities, i.e., whether they can be updated with evidence using the Bayesian conditionalization (6.1).

In fact, there seems to be a straightforward method for doing this, which will be discussed below. It will be shown that the method for updating function P in (6.15) with evidence will follow rule (6.1).

The overall prior probability of the evidence e is, according to (6.15),

$$\sum_i \hat{P}(\lambda_i) \cdot P^i(e) \tag{6.19}$$

and the probability of "λ_i and e" is

$$\hat{P}(\lambda_i) \cdot P^i(e) \tag{6.20}$$

according to (6.14). Hence, the conditional probability of λ_i under e can be stated as

$$\hat{P}(\lambda_i|e) = \hat{P}_e(\lambda_i) = \frac{\hat{P}(\lambda_i) \cdot P^i(e)}{\sum_j \hat{P}(\lambda_j) \cdot P^j(e)}. \tag{6.21}$$

This function is the posterior second-order probability function on inductive methods, i.e. the second-order probability function which is updated on evidence e by using Bayesian conditionalization.[7]

[7] Observe that (6.21) can be read as an application of Bayes's theorem

$$P(H|E) = \frac{P(H) \cdot P(E|H)}{P(E)} \tag{6.22}$$

It should hold that $\sum_i \hat{P}(\lambda_i|e) = 1$, i.e., no evidence should make the sum of second-order probabilities of inductive methods differ from one. It is easily proved that this is indeed the case:

$$\sum_i \hat{P}(\lambda_i|e) = \sum_i \frac{\hat{P}(\lambda_i) \cdot P^i(e)}{\sum_j \hat{P}(\lambda_j) \cdot P^j(e)} = 1. \tag{6.23}$$

Let us proceed to examine how hypotheses are updated when the above updating of the probabilities of inductive methods is included in the updating procedure. The first candidate for a definition of conditional probability with updating of the second-order probabilities will be considered below.

The conditional probability of a hypothesis h, denoted by $P(h|e)$, can be defined by first updating the probabilities of the methods under evidence e using formula (6.21) and then calculating the probability of the hypothesis h using each corresponding method:

$$P^u(h|e) = \sum_i \hat{P}(\lambda_i|e) \cdot P^i(h|e) \tag{6.24}$$

and $P^u(h) = P(h)$ where P is as above.

Observe that the ordinary conditionalization rule naturally holds for first-order probability distributions determined by the inductive methods:

$$P^i(h|e) = P^i(h\&e)/P^i(e). \tag{6.25}$$

Another way of expressing the conditional probability with updating of second-order probabilities would be to import conditioning on e into the numerator and denominator of (6.16):

$$\frac{\sum_i \hat{P}(\lambda_i|e) \cdot P^i(h\&e|e)}{\sum_i \hat{P}(\lambda_i|e) \cdot P^i(e|e)}. \tag{6.26}$$

Note that $P^i(h\&e)$ and $P^i(e)$ in (6.16) need to be replaced by $P^i(h\&e|e)$ and $P^i(e|e)$, respectively, because $P(\lambda_i)$ is also conditioned on e – in other words, it is assumed that e holds.

Since $P(e|e) = 1$, $P(h\&e|e) = P(h|e)$ and according to (6.23), $\sum_i \hat{P}(\lambda_i|e) = 1$, (6.26) equals (6.24) and thus these two ways of defining conditional probability with updating of the second-order probabilities are in fact equivalent.

---

if one sets $H = \lambda_i$, $P(H) = \hat{P}(\lambda_i)$, $P(E|H) = P^i(e)$ and $P(E) = \sum_j \hat{P}(\lambda_j) \cdot P^j(e)$.

It follows from (6.24) that

$$P^u(h|e) = \sum_i \hat{P}(\lambda_i|e) \cdot \frac{P^i(h\&e)}{P^i(e)}. \tag{6.27}$$

According to (6.21), this can be written as

$$\frac{\sum_i \hat{P}(\lambda_i) \cdot P^i(e) \cdot \frac{P^i(h\&e)}{P^i(e)}}{\sum_j \hat{P}(\lambda_j) \cdot P^j(e)}. \tag{6.28}$$

Hence,

$$P^u(h|e) = \frac{\sum_i \hat{P}(\lambda_i) \cdot P^i(h\&e)}{\sum_j \hat{P}(\lambda_j) \cdot P^j(e)}, \tag{6.29}$$

and with (6.15),

$$P^u(h|e) = \frac{P(h\&e)}{P(e)} = \frac{P^u(h\&e)}{P^u(e)} = P(h|e). \tag{6.30}$$

This shows that the updating method with second-order probabilities is in accordance with the ordinary conditionalization rule.[8] However, the striking thing to realize is that (6.29) is exactly the same formula as (6.16), in which the updating of second-order probabilities is not even considered. This indicates that formula (6.24) for updating second-order probabilities must be a kind of quasi-updating, i.e. that updating second order probabilities has no effect on the final probability, which means that there is not much point in introducing the whole idea of updating second-order probabilities, at least not in this way.

This attempt to define an updating procedure for second-order probabilities simply reduces to using prior second-order probabilities. It follows that this kind

[8]It is interesting to note that if one chose not to update the methods themselves with evidence, i.e., that the probabilities of hypotheses were given simply by

$$P(h|e) = \sum_i \hat{P}(\lambda_i) \cdot P^i(h|e) \tag{6.31}$$

instead of (6.24), the result (6.30) would not be achieved. If (6.31) replaces (6.24) in the definition of conditional probability of a hypothesis, the resulting conditional probability must obviously change accordingly at least sometimes since in general $\hat{P}(\lambda_i|e) \neq \hat{P}(\lambda_i)$, see (6.21). Hence no equality which involves (6.24) can hold when (6.24) is replaced by (6.31). This means that (6.31) is not in accordance with the conditionalization rule.

of introduction of second-order probabilities of inductive methods does not solve the original problem, which was dependence on *a priori* assumptions – the choice of an inductive method – in determining probabilities. It seems that the only way second-order probabilities could help in choosing the most rational inductive method is the possibility that one could establish less arbitrary criteria for constraining the choice of the second-order priors than the first-order ones. However, as it stands, such criteria are not in sight.

Chapter 7

Updating the inductive method

This chapter will present a non-Bayesian rule for updating the inductive method on the basis of evidence. It will be argued that the rule should be favoured over the methods of the λ-continuum and that it at least tentatively provides a proposal for solving one of the forms of the problem of induction manifested in inductive logic, namely that concerning the choice of the optimum inductive method in terms of the success criterion for inductive methods as formulated by Carnap (1952).

The rule for updating the inductive method has certain consequences in the constructive interpretation of extendible probability, which was defined in chapter 5. According to extendible probability, truth in infinity means truth in consecutive finite state descriptions. In other words, when more knowledge about the actual world is obtained, S remains true in the light of this evidence. Updating the probability measure in the course of obtaining the evidence has an effect on the probability which corresponds to truth in infinity in the sense of extendible truth, that is, extendible probability. This question will be touched briefly in chapter 8 below, where it will be shown that scientific hypotheses can assume non-zero probabilities when constructive asymptotic probability is combined with the updating rule. This establishes a link between constructive semantics and probability results with the updating rule.

The idea of changing the prior probability distribution with a non-Bayesian rule is not new. For example, Douven (2000) discusses empirical testing of inductive logics. However, more relevant for the present study is the limit-process already discussed in Kuipers (1986, p. 39). The correction rule which will be introduced in (7.23) essentially updates the method analogously to Kuipers's one-step version of the limit process (cf. Kuipers 1986, p. 43). However, despite being able to make some preliminary remarks, Kuipers (1986) does not succeed in evaluating the performance of the one-step system in a far-reaching way, as he clearly recognizes himself.

In this chapter, a method of evaluating the correction rule will be introduced.

The superiority of the correction rule over the methods of the λ-continuum will be argued for.

The proofs in this chapter may contain instances of classical inference, but it is presumed that, should this be the case, the methods can be replaced by their constructive equivalents.

7.1 Inductive skepticism

Carnap regards the extreme method $c^\dagger$, i.e., $\lambda = \infty$ as seemingly inappropriate for sound scientific reasoning on the grounds that it gives no consideration to experience in making expectations or estimations, as long as the experience does not concern the individual mentioned in the hypothesis (e.g., 1962, p. 564; 1952, p. 38).[1] For example, the evidence of n black ravens does not affect the $c^\dagger$-probability that the $n+1$'th raven is black.

However, Carnap's view seems to presuppose that inductive reasoning should be considered sound. Inductive skepticism, on the other hand, suggests that there are no rational grounds for preferring methods which base predictions on observations to any degree. It follows from this that there are no rational grounds for rejecting $c^\dagger$, the anti-inductivist method.

However, it does not follow that $c^\dagger$ should be preferred over the other methods. It is true that $c^\dagger$ does not take experience into account, and thus represents the anti-inductivist attitude, but committing to $c^\dagger$ is too strong a statement on the basis of mere inductive skepticism. Method $c^\dagger$ is optimal only in an extremely heterogeneous universe in which the individuals are evenly distributed among the Q-predicates. Since it is not possible to know whether the universe is constructed like this, one cannot know if $c^\dagger$ is the right choice.

Carnap puts forward the fact that all estimation methods based on the corresponding inductive methods are *self-correcting,* with the sole exception of $c^\dagger$ (1952, p. 63; also p. 44). Suppose that the limit of a relative frequency of a predicate is r in an infinite sequence of individuals. The inductive method λ' is self-correcting if the estimates of the relative frequency of this predicate based on λ' approach r in the limit. Hence, if there is a limiting relative frequency for a particular predicate, all methods except $c^\dagger$ converge toward r.

Let us discuss why the self-correcting methods should be preferred over $c^\dagger$.

[1] In finite domains the quantifiers are defined by means of finite conjunctions of their instances (cf. Carnap 1962, pp. 60-62). This means that if the hypothesis contains a quantifier, a body of evidence concerning any individual is automatically mentioned, which explains why the probability of a universal or existential quantification can change even when using $c^\dagger$ in finite systems. In the infinite case, quantifiers are not conjunctions or disjunctions, but it is unnecessary to discuss this case here because the probabilities for the infinite domains are defined by using the finite domains.

Even though the "inductivist" methods are self-correcting, they give better estimates than $c^\dagger$ only if the universe is not maximally heterogeneous. It seems perhaps that a self-correcting method should be chosen because a maximally heterogeneous universe is such an extreme and unlikely case, but this is only an intuitive feeling based on our unconscious presupposition that the universe has some homogeneity. Such presuppositions cannot be used for justifying inductive inference since they essentially beg the question.

There is a simple argument for choosing a self-correcting method resembling the well-known Reichenbachian justification of induction. This argument goes as follows: if there is a limiting relative frequency for a particular predicate, a self-correcting method will approach this frequency asymptotically, whereas the $c^\dagger$ method will not necessarily do so. In other words, if the universe is of the kind in which learning from experience is possible, then only the self-correcting methods will approach the true relative frequencies, whereas $c^\dagger$ will not.

This does not change the fact that a self-correcting method can only perform better than $c^\dagger$ when the universe has some uniformity. If the universe has no uniformity, $c^\dagger$ will provide better estimates than any self-correcting method. Why should one such prefer a method only because it performs better than $c^\dagger$ in a universe with some uniformity, if there is no reason to suppose that the universe actually has some uniformity?

There does, however, seem to be a way out of the apparent impossibility of making a rational choice between the inductive methods. This will be based on a certain kind of self-correcting method, a correction rule which will take evidence into account not only by using the conditionalization rule on the probabilities but also by adjusting the inductive method itself.[2]

It will argued that the correction rule qualifies better than any given inductive method in a set of inductive methods whose complement can be made arbitrarily small (but not empty). This does not mean that the correction rule will necessarily perform better than a particular λ', but if one is to choose between λ' and the correction rule, there is an argument for preferring the correction rule provided that there is no other reason to favour λ'. Since $c^\dagger$ is no more justified than any other constant method, this reasoning also applies to $\lambda' = c^\dagger$. Hence, although there is no reason to prefer one constant method over another, there is a reason to prefer the correction rule over any constant method.

If the correction rule is adopted, the possibility of inductive inference is dependent on the nature of the evidence. If the evidence obtained shows some uniformity, the updated inductive method is adjusted accordingly, making probabilistic

[2]Carnap acknowledges the possibility of changing the method according to its past performance (1952, p. 55); on the other hand, he expresses doubts about such adjustments and prefers *a priori* considerations in choosing the value of λ (1968, pp. 313-14).

inductive inference possible. However, it is not possible to examine in this study whether the above argument in favour of the correction rule really justifies inductive inference.

7.2 Immodesty

This section discusses the *immodesty* criterion, which has been suggested in the literature as a necessary criterion for an acceptable inductive method. The concept was originally introduced by Lewis (1971); comments include at least Spielman (1972), Lewis (1974), Pietarinen (1974), Horwich (1982) and Festa (1994).

An inductive method is immodest if it evaluates its own predictions as the most accurate ones. In other words, the estimates of a certain magnitude M made by a method λ are compared to real values of M in all possible state descriptions. The λ-weighted average of the errors obtained this way is then the performance indicator of λ as evaluated by λ itself.

Immodesty can be considered as a necessary criterion for inductive methods. If an inductive method λ is not the most accurate one when evaluated by itself, it clearly cannot be considered as the optimum one without contradiction.[3]

According to Lewis's original paper, almost no inductive method is immodest. On the other hand, Spielman (1972) arrived at a different conclusion, saying that all inductive methods *are* in fact immodest.

Lewis (1974) explains the deviation of results by different approximations used in the calculations. Lewis himself uses Carnap's own accuracy criterion (1952), namely the mean square error of a method (see section 7.3 below). Carnap's measure, however, involves an approximation which simplifies the calculations. The approximation is valid only if the universe is very large compared to the sample which is used for prediction. According to Lewis, if one replaces this approximation with the exact formulas, one gets Spielman's results saying that all inductive methods are immodest.

However, Lewis does not explicitly show this. Since the motivation for using the approximation is the mathematical intractability of the exact formula, it is unclear whether Spielman's results are as easily obtainable as Lewis claims.

On the other hand, Horwich claims to have proved the immodesty of all inductive methods without having to resort to any approximations concerning the size of the universe (1982, pp. 87-90). Festa has a somewhat similar approach (1994, pp. 41-44).

To sum up this discussion, the final outcome of immodesty considerations is not quite clear, but if Horwich's result is taken to be valid, immodesty does not

[3] In other words, if λ is chosen, it itself suggests that another method should be chosen.

rule any inductive method and thus cannot be a basis for choosing the correct method.

7.3 Carnap's measure of success

In this section, I will elaborate Carnap's original work (1952) on the measure of success or performance of an inductive method, namely the mean square error.

The expression

$$Est_{\lambda,\kappa}(x, X_i) = \frac{X_i + (\frac{1}{\kappa})\lambda}{x + \lambda} \tag{7.1}$$

is the λ-estimate of the relative frequency of property Q_i in a sample of cardinality x obtained from the universe of n individuals w_n (κ being the number of Q-predicates in the language). (Cf. Carnap 1952, p. 33.)

This estimate is determined by using λ on the basis of a sample of x individuals among which X_i have the property Q_i. It is assumed that the size of the whole universe n is very large compared to x so that it holds that the whole domain n approximately represents the unobserved part $U(n, x)$ of the universe, whose cardinality is $n - x$. Function (7.1) is a random function which depends on the random variable X_i. The expected value and variance of X_i can be obtained from x and w_n, in the way which will be described in what follows (cf. Carnap 1952, pp. 57-58). This value will then be used to determine the error of the estimate (7.1) (cf. Carnap 1952, p. 58).

Let r_i be the relative frequency of Q_i in the universe. Because n is very large compared to x, r_i is approximately the relative frequency of Q_i in $U(n, x)$ as well.

The proportion of samples with a given absolute frequency of Q_i, s_i, among all the samples of size x is approximately

$$P(X_i = s_i) \approx \binom{x}{s_i} r_i^{s_i}(1 - r_i)^{x - s_i}, \tag{7.2}$$

in other words, the random variable X_i approximately follows the binomial distribution.

The binomial distribution is customarily used to represent situations in which the random variable denotes the number of successes in an experiment which is repeated x times and where the probability of success remains constant (in this case r_i) in each trial. The critical part of this use of the formula concerns the justification of assuming the probability of success to be constant (i.e., r_i) in each trial. If the sample is large compared to the total population, a collection of observations about individuals in the sample does have an effect on what the

expected relative frequency of Q_i-individuals in the rest of the sample is. Hence, formula (7.2) gives only approximately correct probabilities when n is sufficiently large compared to x. When $n \to \infty$, (7.2) is the value of the corresponding limit.

The expected value (also called the mean value in this study) of a discrete random variable X is defined as

$$E(X) = \sum_k x_k \cdot p_k \tag{7.3}$$

where x_k are the different possible values of X and the p_k are the probabilities $P\{X = x_k\}$, provided that the series converges absolutely. Hence, the mean value of X is the weighted average of X.

Notice that sometimes in the literature the mean value of a magnitude C often signifies simply the average value of C:

$$\sum_{k=1}^{n} c_k \cdot \frac{1}{n} \quad (C = c_1, ..., c_n). \tag{7.4}$$

The variance of a random variable X is defined as

$$D^2(X) = E((X - \mu)^2), \tag{7.5}$$

where $\mu = E(X)$.

For any variable X following the binomial distribution with parameters p and y, one has

$$E(X) = py, \tag{7.6}$$

$$Var(X) = py(1 - p), \tag{7.7}$$

for the expected value of X, where p is the number of repetitions and y is the probability of success in each trial. Hence,

$$E(X_i) = xr_i \tag{7.8}$$

and

$$Var(X_i) = xr_i(1 - r_i). \tag{7.9}$$

The error of the estimate function $Est_{\lambda,\kappa}(x, X_i)$ of the random variable X_i introduced in (7.1) above is also determined on the basis of r_i, provided again that n is sufficiently large compared to x since $Est_{\lambda,\kappa}(x, X_i)$ gives the estimated frequency of Q_i for the part of the domain whose size is $n - x$:

$$Err_{\lambda,\kappa}(x, X_i, r_i) = Est_{\lambda,\kappa}(x, X_i) - r_i. \tag{7.10}$$

The presentation of the results below differs slightly from that of (Carnap 1952, pp. 62-65). By virtue of (7.10), (7.1) and (7.8),

$$E(Err_{\lambda,\kappa}(x, X_i, r_i)) = \frac{xr_i + (\frac{1}{\kappa})\lambda}{x + \lambda} - r_i = \frac{(\frac{1}{\kappa} - r_i)\lambda}{x + \lambda}. \tag{7.11}$$

The following general result holds for variance:

$$Var(aX + b) = a^2 Var(X), \tag{7.12}$$

where a and b are constants. Because the variance of the error (7.10) above is the same as the variance of the estimate, it holds by (7.12) that

$$\begin{aligned} &Var(Err_{\lambda,\kappa}(x, X_i, r_i)) = Var(Est_{\lambda,\kappa}(x, X_i)) = \\ &Var(\frac{X_i + (\frac{1}{\kappa})\lambda}{x + \lambda}) = Var(\frac{X_i}{x + \lambda} + \frac{\lambda}{\kappa(x + \lambda)}) \\ &= \frac{1}{(x + \lambda)^2} Var(X_i) = \frac{xr_i(1 - r_i)}{(x + \lambda)^2}. \end{aligned} \tag{7.13}$$

The following is valid for an arbitrary random variable Y:

$$[E(Y)]^2 + Var(Y) = E(Y^2). \tag{7.14}$$

The mean (in the sense of expected) square error is the sum of square errors for each sample in x multiplied by the probability of the sample, i.e., the weighted average of square errors. Hence, the mean square error $E(Err^2_{\lambda,\kappa}(x, X_i, r_i))$ with respect to Q_i equals

$$[E(Err_{\lambda,\kappa}(x, X_i, r_i))]^2 + Var(Err_{\lambda,\kappa}(x, X_i, r_i)). \tag{7.15}$$

Using the above results, one can derive the formula for the mean square error with respect to Q_i:

$$E(Err^2_{\lambda,\kappa}(x, X_i, r_i)) = \frac{xr_i(1 - r_i) + (\frac{1}{\kappa} - r_i)^2\lambda^2}{(x + \lambda)^2}. \tag{7.16}$$

(Cf. Carnap 1952, p. 65.) Because the above does not depend on X_i, it can be written as

$$\overline{Err}^2_{\lambda,\kappa}(x, r_i). \tag{7.17}$$

With respect to all Q_i, the mean square error is:

$$\frac{1}{\kappa} \sum_{i=1}^{\kappa} \overline{Err}^2_{\lambda,\kappa}(x, r_i). \tag{7.18}$$

This can also be called the average mean square error or the average expected square error. Carnap (1952, p. 65-67) shows that (7.18) equals

$$\frac{x - \frac{\lambda^2}{\kappa} + (\lambda^2 - x)\sum_{i=1}^{\kappa} r_i^2}{\kappa(x+\lambda)^2}, \tag{7.19}$$

which means that the mean square error with respect to all Q-predicates in a particular universe (state description) depends only on the sample size x, the method λ and the sum $\sum_i r_i^2$, which is the degree of order of the state description in question (cf. section 6.3 above). Hence, (7.18) can be denoted by

$$\overline{\forall_Q Err}^2_{\kappa,\lambda}(x, \sum_i r_i^2). \tag{7.20}$$

7.4 The correction rule

The idea of the *correction rule* to be introduced in this section is to update the method currently employed in the course of obtaining more and more information about the actual (unknown) state description. It is hoped that this would result in a better performance than any constant inductive method. This prospect will be examined in section 7.5 below.

We saw above in chapter 6 that there is an optimum value of λ for any given state description, which can be calculated from the degree of order of the state description. The formula will be repeated here using slightly different formalism.

The optimum method for a given degree of order $d_o(w)$, i.e., that which minimizes the mean square error for this degree of order, is

$$\delta^{-1}(d_o(w)) = \frac{1 - d_o(w)}{d_o(w) - \frac{1}{\kappa}}. \tag{7.21}$$

The problem with Carnap's measure of success is that it can be calculated only for given degrees of order. Hence, the optimum method can also be determined only for given degrees of order. Consider the question of which inductive method is the optimum one for the actual state description. One usually knows only a relatively small part of the universe; it is just this fact that makes the use of inductive methods necessary. Therefore, the degree of order of the universe is usually not known, which means that there is no way to find the optimum inductive method. (Cf. Carnap 1952, p. 71.) No method seems to be excluded on the basis of the mean square error performance criterion. In particular, the anti-inductivist method $c^\dagger$ cannot be rejected.

However, the situation may change with a non-Bayesian updating rule like the correction rule below. Possibly the correction rule will yield better estimates measured by the mean square error than any of the non-optimum constant methods.

If the universe of discourse has n individuals and, according to the state description w_n, all of them satisfy a certain Q-predicate, w_n's degree of order is maximal. On the other hand, if the individuals are distributed evenly among all possible Q-predicates, w_n's degree of order is minimal. Other cases fall between these two extremes. Hence, since each description of a sample of individuals can be considered as a state description in a universe of the size of the sample, there is an optimum method for each given sample.

The most trivial requirement for any correction rule is that it has to yield the optimum method when the sample size consists of all individuals in the domain. For finite domains, it is easy to construct a rule that fulfils this requirement. The following line captures the main idea of such a rule:

$$Corr(w_n, \lambda_a) = \delta^{-1}(d_o(w_n)), \tag{7.22}$$

where $Corr(w_n, \lambda_a)$ is the method when the data consists of a state description w_n in the sample of n individuals, λ_a is the initial method and $\delta^{-1}(d_o(w_n))$ the optimum method for the state description w_n. Hence, if the sample w_n is the whole state description, (7.22) yields the optimum method for that state description.

Rule (7.22) is an extreme one in the sense that it simply changes the inductive method to comply with the evidence received, which means that presuppositions about the best method can play no role there.

Following this idea, a correction rule for an arbitrary sample can be defined simply as follows:

$$\begin{cases} \Theta(\lambda_a, x) = \lambda_a, \text{if } x = 0 \\ \Theta(\lambda_a, x, w_x) = \delta^{-1}(d_o(w_x)), \text{if } x > 0 \text{ and } d_o(w_x) > \frac{1}{\kappa}, \\ \Theta(\lambda_a, x, w_x) = \Theta(\lambda_a, x, w_{x-1}), \text{if } x > 0 \text{ and } d_o(w_x) = \frac{1}{\kappa}. \end{cases} \tag{7.23}$$

The Θ-function in (7.23) updates the method in each w_x to correspond to the optimum method for that state description (starting with the initial method λ_a). When more information is received, i.e., x increases, the method is updated correspondingly.

The justification for the last condition above is that when the sample is maximally heterogeneous, no inductive method can react upon any uniformity in the sample and hence every method will provide the same estimate for all Q_i, namely $\frac{1}{\kappa}$.[4]

A more general form of the rule can be achieved by modifying it in two respects: 1) adding a condition for a lower limit of cardinality x when the rule may start to be applied (this will turn out to be essential in the final conclusions about the performance of the rule in section 7.5.3), 2) adding a caution factor which

[4] I am grateful to Theo Kuipers for this observation.

tells us how much the inductive method should be adjusted at each step (cf. section 7.5.3.7).

7.5 Performance

One can now ask whether the correction rule Θ outranks the non-optimum constant inductive methods with respect to the mean square error.

As the Θ-rule updates the method according to evidence, the mean square error of the correction rule cannot be calculated by treating the inductive method as a constant, which means that the calculation is more difficult to carry out than in Carnap (1952).

The task is to show that the mean square error of Θ is smaller than that of a given constant method when the sample is large enough.

It is evident that the expected degree of order of the sample approaches that of the universe as the size of the sample increases. This means that the expected value of Θ will be closer to the optimum method than a given constant method when the size of the sample is large enough. Carnap proves that the mean square error produced by λ decreases when λ approaches λ^{Δ} (1952, pp. 68-69). However, the mean square error of the Θ-rule does not necessarily equal the mean square error of the expected value of Θ, which makes the issue technically complicated.

Since samples whose degree of order is close to that of the universe become more probable when the sample size increases, the weight of this kind of sample also increases in determining the mean square error. Hence, even without being able to derive the mean square error of Θ, one may be able to show that it must be smaller than that of a given constant method under certain circumstances.

But even though one could show that Θ qualifies better than a given constant λ when x is large enough, the estimates for smaller samples matter as well. In other words, one may obtain a significant number of larger mean square errors with Θ than with a given constant method before a certain sample size x. It is not clear that choosing Θ at the outset is the optimal solution.

The situation is analogous when comparing constant inductive methods. The difference between the mean square errors of two self-correcting constant methods converge to zero in most cases when the sample size increases (as will be shown in section 7.5.2). Some commentators have argued that for this reason the choice of the inductive method does not matter, but this is not correct. The choice of the method does matter when one is not acting in the "asymptotic limit" but in a more immediate world when the methods do produce different errors. If one wishes to obtain a correct estimate, which inductive method one uses is not insignificant.

Observe that it is not only the difference between single estimates by two constant methods which is at stake here. A non-optimal method gives rise to a

number of more inaccurate estimates than the optimum one when the sample size grows.

Similarly, one cannot straightforwardly prefer the Θ-rule just because it will eventually produce a smaller error than a given constant method. It is possible that the Θ-rule produces a large number of very inaccurate estimates compared to those of a given constant λ. It is even possible that this property of Θ makes it knowably more inaccurate in an overall evaluation than most constant methods. If a method is knowably more inaccurate than most other methods and knowably more accurate than a few methods, pure guesswork in choosing the inductive method may result in a statistically better result than the Θ-rule.

A more general measure of performance than the mean square error is obtained when one considers a series of nested samples, where new individuals are added to the sample previously obtained. For example, one first determines the error by using a set of variables X_i for the sample size x, and for the sample size $x+1$ one adds a random variable representing the value of the $x+1$'th individual to the values of the variables X_i already obtained.[5]

The reason for considering a series of samples is to find the optimum estimation method for the whole process of obtaining information from a population. Even if there is no criterion for choosing the optimum method for an individual prediction, perhaps there is one for a series of predictions. The formulation of a performance measure with nested samples thus has some significance.

The *cumulative square error* represents the total error one makes in consecutive estimates when new individuals are added to the same sample. It will be shown below in section 7.5.4 that the *mean cumulative square error* can be defined as the mean (expected) value of all possible cumulative square errors, or, equivalently, as the sum of mean square errors for consecutive samples.

However, it will turn out in section 7.5.3 that there is another way of evaluating the performance of a more general form of the Θ-rule which includes the restric-

[5]It helps to understand the situation when one observes that the value of X_i can be calculated by consecutive answers to the questions such as 'is the 1. individual in the sample Q_i?', 'is the 2. individual in the sample Q_i?' etc. Hence, X_i can be represented by means of a sum of indicator variables $1_{i(1)}, 1_{i(2)}, ..., 1_{i(x)}$, where $1_{i(y)}$ is 1 if the y'th individual is Q_i and otherwise $1_{i(y)} = 0$:

$$X_i = \sum_{y=1}^{x} 1_{i(y)}. \tag{7.24}$$

The random variable for the sample size $x+1 \leq n$ can be expressed as

$$\sum_{y=1}^{x+1} 1_{i(y)} = \sum_{y=1}^{x} 1_{i(y)} + 1_{i(x+1)} = X_i + 1_{i(x+1)}. \tag{7.25}$$

Hence, the value of this variable is not independent of the value of X_i.

tion that the rule may start to be applied only after the sample is large enough, see point 1) on p. 110 above, without resorting to a performance measure using nested samples.

7.5.1 Convergence of mean square error with constant methods

The first steps toward the evaluation of the correction rule as compared to constant inductive methods will be taken in this section by introducing some findings on the convergence of the mean square error.

Let us first examine the situation for $\lambda < \infty$. The mean square error (7.19) can be written as

$$\frac{x - \frac{\lambda^2}{\kappa} + \lambda^2 \sum r_i^2 - x \sum r_i^2}{\kappa x^2 + 2\kappa\lambda x + \kappa\lambda^2}. \tag{7.26}$$

For large values of x, the denominator of (7.26) is approximately equal to x^2. Consider first the case in which $\sum r_i^2 < 1$. Since (7.26) is then approximately equal to $\frac{1}{x}$ for a sufficiently large x, its convergence rate can be compared to that of $\frac{1}{x}$.

Consider then the case in which $\sum r_i^2 = 1$. Now (7.26) reduces to a form in which x^2 occurs in the denominator. If $\lambda = 0$, the mean square error is zero in this case (cf. Carnap 1952, p. 69).

Then consider the case when $\lambda = \infty$. The limit convention of Carnap (1952, p. 33) means that for any function $f(\lambda)$ the value of $f(\lambda)$, when $\lambda = \infty$, is $\lim_{\lambda \to \infty} f(\lambda)$. Hence, to achieve the mean square error of λ for a particular x, one must consider the limit of (7.27) below.

Beside (7.26), the mean square error (7.19) can also be written as

$$\begin{gathered}\frac{x - \frac{\lambda^2}{\kappa} + \lambda^2 \sum r_i^2 - x \sum r_i^2}{\kappa\lambda^2(\frac{x^2}{\lambda^2} + \frac{2x}{\lambda} + 1)} = \\ \frac{\frac{x}{\kappa\lambda^2} - \frac{1}{\kappa^2} + \frac{\sum r_i^2}{\kappa} - \frac{x \sum r_i^2}{\kappa\lambda^2}}{\frac{x^2}{\lambda^2} + \frac{2x}{\lambda} + 1}.\end{gathered} \tag{7.27}$$

Consider the last form above when $\lambda \to \infty$. The first two terms in the denominator clearly tend to zero, which entails that the denominator tends to 1. The first and last terms of the numerator tend to zero as well. Hence, the whole expression tends to

$$\frac{\sum r_i^2}{\kappa} - \frac{1}{\kappa^2}. \tag{7.28}$$

This constant is the mean square error for $\lambda = \infty$.[6]

If $\sum r_i^2 > \frac{1}{\kappa}$, (7.28) is greater than zero. The remaining case is $\sum r_i^2 = \frac{1}{\kappa}$, in which (7.28) is zero (as Carnap 1952, p. 69 also proves).

7.5.2 Comparing two constant methods

7.5.2.1 The difference between mean square errors

In this section it will be proved that the difference between mean square errors of two arbitrary constant methods converges to zero with the rate proportional to $\frac{1}{x^2}$ in most cases.

Let us first examine the case in which $\lambda_1, \lambda_2 < \infty$.

If $\sum r_i^2 = 1$ and $\lambda_1 = 0$ and $\lambda_2 > 0$, the mean square error with λ_1 is 0 and converges to zero at a rate proportional to $\frac{1}{x^2}$ with λ_2; hence, their difference converges to zero at a rate proportional to $\frac{1}{x^2}$.

If $\sum r_i^2 = 1$ and $\lambda_1, \lambda_2 > 0$, then the mean square errors of both λ_1 and λ_2 converge to zero at a rate proportional to $\frac{1}{x^2}$, which means that the difference between them converges at least at this rate.

Then assume that $\sum r_i^2 < 1$. Consider the expression (7.26) for the mean square error when $x \to \infty$. One needs to find out the rate of convergence of

$$\frac{x - \frac{\lambda_1^2}{\kappa} + \lambda_1^2 \sum r_i^2 - x \sum r_i^2}{\kappa(\lambda_1 + x)^2} - \frac{x - \frac{\lambda_2^2}{\kappa} + \lambda_2^2 \sum r_i^2 - x \sum r_i^2}{\kappa(\lambda_2 + x)^2}. \tag{7.29}$$

It is obvious that the two middle terms in the numerators of the two mean square errors in (7.29), of the form $\frac{\lambda^2}{\kappa}$ and $\lambda^2 \sum r_i^2$, converge to zero at a rate which is proportional to x^2.

It is also obvious that the difference

$$\left| \frac{\frac{\lambda_1^2}{\kappa} + \lambda_1^2 \sum r_i^2}{\kappa(\lambda_1 + x)^2} - \frac{\frac{\lambda_2^2}{\kappa} + \lambda_2^2 \sum r_i^2}{\kappa(\lambda_2 + x)^2} \right| \tag{7.30}$$

is a decreasing function of x.

The remaining question is how fast the difference

$$\begin{aligned} &\frac{1}{\kappa x}(1 - \sum_i r_i^2)\frac{x^2}{(\lambda_1 + x)^2} - \frac{1}{\kappa x}(1 - \sum_i r_i^2)\frac{x^2}{(\lambda_2 + x)^2} \\ &\quad = \frac{1}{\kappa x}(1 - \sum_i r_i^2)[\frac{x^2}{(\lambda_1 + x)^2} - \frac{x^2}{(\lambda_2 + x)^2}] \end{aligned} \tag{7.31}$$

[6]Since the square error of $\lambda = \infty$ is the same for every distribution of Q-predicates in the sample, the result (7.28) can also be obtained directly from the average square error denoted by the expression (7.58) below.

converges.

The latter factor above becomes

$$\begin{aligned}\frac{x^2}{(\lambda_1+x)^2}-\frac{x^2}{(\lambda_2+x)^2} =& \frac{x^2(\lambda_2+x)^2-x^2(\lambda_1+x)^2}{(\lambda_1+x)^2(\lambda_2+x)^2}\\ =& \frac{x^2(\lambda_2^2+2\lambda_2 x+x^2-\lambda_1^2-2\lambda_1 x-x^2)}{(\lambda_1+x)^2(\lambda_2+x)^2}\\ \leq& \frac{x^2(\lambda_2^2-\lambda_1^2+2x(\lambda_2-\lambda_1))}{x^4}.\end{aligned} \tag{7.32}$$

The expression $\frac{\lambda_2^2-\lambda_1^2}{x^2}$ clearly converges to zero with a rate proportional to $\frac{1}{x^2}$. The expression

$$\frac{2}{x}(\lambda_2-\lambda_1) \tag{7.33}$$

converges in proportion to $\frac{1}{x}$, which means that (7.31) converges to zero with the speed proportional to $\frac{1}{x^2}$. Since the denominator of (7.32) increases with x, it is also clear that the absolute value of this part of the difference between mean square errors of two methods is a decreasing function of x.

The remaining case is when $\lambda_1=\infty$ and $\lambda_2<\infty$. Then the mean square error of λ_1 is (7.28).

If $\sum_i r_i^2=\frac{1}{\kappa}$, (7.28) is 0 with λ_1 and the mean square error of λ_2 converges in proportion to $\frac{1}{x}$ (see section 7.5.1 above). Hence, the difference between λ_1 and λ_2 is converges to zero in proportion to $\frac{1}{x}$.

If $\sum_i r_i^2>\frac{1}{\kappa}$, by (7.28) the mean square error of $\lambda_1=\infty$ is a positive constant which can be denoted by c. The mean square error of $\lambda_2<\infty$ is proportional to $\frac{1}{x}$. Consider then the difference between the mean square error of λ_2 and λ_1, which is approximately $c-\frac{1}{x}$ for large values of x. Since $\frac{1}{x}$ tends to zero, the difference converges to c with the speed proportional to $\frac{1}{x}$.

7.5.3 Comparing the Θ-rule and non-optimum constant methods

Recall that our purpose is to evaluate the performance of Θ compared to a non-optimum inductive method.

The easiest case for comparing the performance of Θ and a non-optimum constant method λ is when $\sum r_i^2=1$. In this case every sample must be uniform, hence the value of Θ must be 1 after observing a sample of at least one individual. The choice of the prior method λ_a in (7.23) matters only for the estimates prior to any evidence and the purpose of the Θ-rule is not to provide guidance for

the *a priori* selection of a constant method. Hence, although the given method λ can perform better for the estimate prior to evidence than λ_a, this is not relevant for evaluating the performance of Θ. In situations where at least one individual has been observed, the Θ-rule yields the optimum method $\lambda = 0$ and thus its performance is equal to that of the optimum method.

However, the situation is more complicated if $\sum r_i^2 < 1$. This case will be analysed in the following sections.

7.5.3.1 The expected degree of order

Some preliminary derivations will be useful.

One obtains the expected degree of order of a sample w_x if one assumes a particular degree of order for the whole universe and the sample is selected by means of random sampling from the population of individuals in the whole universe.

The expected degree of order of the sample w_x is

$$\begin{aligned} E(d_o(w_x)) = E(\sum_i [R(w_x, i)]^2) = \\ \sum_i E(\frac{X_i^2}{x^2}) = \frac{1}{x^2} \sum_i E(X_i^2). \end{aligned} \tag{7.34}$$

In the above, w_x is a sample and i goes through all the Q-predicates. Because

$$E(X_i^2) = Var(X_i) + (E(X_i))^2, \tag{7.35}$$

one obtains

$$E(X_i^2) = xr_i(1 - r_i) + r_i^2 x^2, \tag{7.36}$$

using (7.8) and (7.9), where r_i is the relative frequency of Q_i in the whole universe. Hence, by (7.34),

$$\begin{aligned} E(d_o(w_x)) &= \frac{1}{x^2} \sum_i [xr_i(1 - r_i) + r_i^2 x^2] \\ &= \frac{1}{x} \sum_i r_i + (1 - \frac{1}{x}) \sum_i r_i^2 \\ &= \frac{1}{x} + (1 - \frac{1}{x}) \sum_i r_i^2 \\ &= \sum_i r_i^2 + \frac{1}{x}(1 - \sum_i r_i^2) \geq \sum_i r_i^2. \end{aligned} \tag{7.37}$$

It is interesting to note that the degree of order of the whole universe can be stated as the inverse function of the expected degree of order of the sample w_x:

$$\sum_i r_i^2 = \frac{E(d_o(w_x)) - \frac{1}{x}}{1 - \frac{1}{x}}. \tag{7.38}$$

7.5.3.2 Improbable samples

Recall the definition (7.23) of the Θ-rule. The Θ-rule yields a method which is the optimum one on the basis of the sample obtained. Hence, if the sample tends to a particular degree of order when the sample size grows without an upper bound, the Θ-rule obviously tends to the optimum method of this degree of order.

Consider a sample size x in which

$$|\sum_i r_i^2 - E(d_o(w_x))| < |\sum_i r_i^2 - d_1|, \tag{7.39}$$

for a given degree of order d_1 ($\frac{1}{\kappa} \leq d_1 \leq 1$). This condition means that $E(d_o(w_x))$ is closer to the degree of order of the whole universe than d_1. When the condition is fulfilled for some x, it is clear from (7.37) that it will be fulfilled for all samples larger than x as well since $E(d_o(w_x))$ will be closer to $\sum_i r_i^2$ when x increases. It is also clear that there is such x that (7.39) holds for an arbitrary $d_1 \neq \sum_i r_i^2$.

The performance of Θ in terms of mean square error will be compared below to that of λ',

$$\lambda' = \delta^{-1}(d_1). \tag{7.40}$$

Let ϵ be such that if $d_1 > \sum_i r_i^2$, it holds that

$$(\exists x)(d_1 - E(d_o(w_x))) \geq \epsilon) \tag{7.41}$$

and if $d_1 < \sum_i r_i^2$, it holds that

$$(\exists x)(\sum_i r_i^2 + (\sum_i r_i^2 - d_1) - E(d_o(w_x)) \geq \epsilon). \tag{7.42}$$

If $d_1 > \sum_i r_i^2$, it follows from (7.41) that $|\sum_i r_i^2 - d_1| \geq \epsilon$. Consider then the case $d_1 < \sum_i r_i^2$. Since $E(d_o(w_x)) \geq \sum_i r_i^2$, it follows now from (7.42) that $|\sum_i r_i^2 - d_1| \geq \epsilon$.

The idea behind ϵ in (7.41) and (7.42) is the following. Consider a degree of order d_2 for which $|\sum_i r_i^2 - d_2| \geq |\sum_i r_i^2 - d_1|$. The distance of d_2 from $\sum_i r_i^2$ is then obviously at least ϵ. Suppose first that $d_2 < \sum_i r_i^2$. Because $E(d_o(w_x)) \geq \sum_i r_i^2$, it follows that $|d_2 - E(d_o(w_x))| \geq \epsilon$ holds. Suppose then that $d_2 > \sum_i r_i^2$.

If $d_1 > \sum_i r_i^2$, then (7.41) obviously entails $|d_2 - E(d_o(w_x))| \geq \epsilon$. On the other hand, if $d_1 < \sum_i r_i^2$, then $d_2 \geq \sum_i r_i^2 + (\sum_i r_i^2 - d_1)$, which together with (7.42) entails $d_2 - E(d_o(w_x))) \geq \epsilon$ for the same x for which (7.42) holds.

Hence, if $|\sum_i r_i^2 - d_2| \geq |\sum_i r_i^2 - d_1|$ for some d_2, the distance of d_2 from $E(d_o(w_x))$ is at least ϵ for some x (and thus clearly for all $y > x$ because $E(d_o(w_x))$ approaches $\sum_i r_i^2$ when x increases).[7]

Consider now the probability

$$P\{|d_o(w_x) - E(d_o(w_x))| \geq \epsilon\}. \tag{7.43}$$

Chebychev's Inequality says that

$$P\{|d_o(w_x) - E(d_o(w_x))| \geq b\sqrt{Var(d_o(w_x))}\} \leq \frac{1}{b^2} \tag{7.44}$$

for all $b > 0$ and all x (the precondition $\sqrt{Var(d_o(w_x))} > 0$ is clearly satisfied). As probability (7.43) is under consideration here, it is important to observe that Chebychev's Inequality also holds trivially for any b for which $b\dot{V}ar(d_o(w_x)) = \epsilon$.

It will be proved below in 7.5.3.6 that $Var(d_o(w_x))$ converges to zero at least at a speed proportional to $\frac{1}{x}$ when $x \to \infty$.

Let as choose b so that $b\sqrt{Var(d_o(w_x))} = \epsilon$ is satisfied:

$$b = \frac{\epsilon}{\sqrt{Var(d_o(w_x))}}. \tag{7.45}$$

Then b^2 increases at not less than the same rate as $Var(d_o(w_x))$ converges, i.e., at least proportionally to x.

By (7.44), (7.43) has to converge at a rate which is proportional to $\frac{1}{b^2}$. From the choice of b above, it then follows that (7.43) has to converge at least at a speed which is proportional to $\frac{1}{x}$.

It follows that those values of $d_o(w_x)$ whose distance from $E(d_o(w_x))$ is at least ϵ, i.e., (7.43) holds, have a vanishing probability when the sample size grows. To be more precise, their combined probability diminishes at a rate which is at least proportional to $\frac{1}{x}$.

7.5.3.3 Comparing errors of two types of method

Consider now the sample w_x, whose degree of order $d_o(w_x)$ satisfies

$$|d_o(w_x) - E(d_o(w_x))| \geq \epsilon. \tag{7.46}$$

[7] Note that for some d_3, $|d_3 - E(d_o(w_x))| \geq \epsilon$ can be true even if $|\sum_i r_i^2 - d_3| < |\sum_i r_i^2 - E(d_o(w_x))|$, but this does not matter for the present argument.

It was shown in section 7.5.3.2 that the probability of this kind of sample converges to zero at a rate proportional to $\frac{1}{x}$.

Let us then consider degrees of order whose distance from $E(d_o(w_x))$ is less than ϵ, i.e., the set

$$G_\epsilon(x) = \{w_x \big| |d_o(w_x) - E(d_o(w_x))| < \epsilon\}. \tag{7.47}$$

If $w_x \in G_\epsilon(x)$, the value of $\Theta(\lambda_a, x, w_x)$ is closer to $E(d_o(w_x))$ than λ', and thus by (7.39) also closer to the optimum method corresponding to the degree of order of the universe.

As mentioned above on p. 111, the mean square error of the given method λ decreases when it approaches the optimum method λ^Δ. If $\lambda^\Delta = \infty$ holds, the mean square error is clearly a decreasing function of the distance between λ and λ^Δ. If $0 < \lambda^\Delta < \infty$ holds,[8] the mean square error decreases when λ approaches λ^Δ either from below or from above, but it is not immediately obvious that the error is a decreasing function of the distance between λ and λ^Δ, since the rate of decrease may be different depending on which side λ^Δ is approached from.

Hence, when for some λ'' it holds that $\lambda' < \lambda^\Delta < \lambda''$ or $\lambda'' < \lambda^\Delta < \lambda'$, λ' can be such that even if $|\lambda'' - \lambda^\Delta| < |\lambda' - \lambda^\Delta|$, the mean square error of λ'' is greater than that of λ' in some x. In such a case, the constant ϵ in (7.41) will be chosen so that d_1 can be replaced by $\delta(\lambda^o)$, satisfying the following constraint for all λ: if $|\lambda - \lambda^\Delta| < |\lambda^o - \lambda^\Delta|$, the mean square error of λ is smaller than that of λ'. Since the mean square error converges proportionally to $\frac{1}{x}$ for all $\lambda < \infty$ on each side of λ^Δ, it is presumed here that λ^o can be considered roughly constant for each x; the same follows trivially in the case $\lambda' = \infty$ because the mean square error of such non-optimum λ' does not converge to zero.

Because this replacement does not change the results obtained for λ' in this discussion, we will adhere to the constant d_1 in the derivations below. It can therefore be said that for w_x in $G_\epsilon(x)$, the corresponding value of Θ produces a smaller mean square error than λ'.

The value of $\delta(\Theta)$ belongs to $G_\epsilon(x)$ with a probability which is approximately $1 - \frac{1}{x}$ for large x. Let us compare the *expected* mean square error of Θ with that of λ'. In the derivations below, the values of $\Theta < \infty$ for which $\delta(\Theta) \notin G_\epsilon(a')$ will be denoted by $\Theta_1(a')$ and the rest of the values by $\Theta_2(a')$.

If the inequality

$$\begin{aligned} &\frac{1}{a'}(\overline{\forall_Q Err}^2_{\kappa,\lambda'}(x, \sum_i r_i^2) - \overline{\forall_Q Err}^2_{\kappa,\Theta_1(a')}(x, \sum_i r_i^2)) + \\ &(1 - \frac{1}{a'})(\overline{\forall_Q Err}^2_{\kappa,\lambda'}(x, \sum_i r_i^2) - \overline{\forall_Q Err}^2_{\kappa,\Theta_2(a')}(x, \sum_i r_i^2)) > 0 \end{aligned} \tag{7.48}$$

[8] $\lambda^\Delta = 0$ was discussed at the beginning of section 7.5.3.

holds for some a', some cardinality x and each value of $\Theta_1(a')$ and $\Theta_2(a')$ satisfying the above requirements, the expected difference between the mean square errors of λ' and Θ in x is greater than zero. This means that the expected mean square error of Θ is smaller than that of λ' in x.

The expression (7.48) can be written as

$$\begin{aligned}&\overline{\forall_Q Err}^2_{\kappa,\lambda'}(x, \sum_i r_i^2) - \overline{\forall_Q Err}^2_{\kappa,\Theta_2(a')}(x, \sum_i r_i^2)) > \\ &\frac{1}{a'}(\overline{\forall_Q Err}^2_{\kappa,\Theta_1(a')}(x, \sum_i r_i^2)) - \overline{\forall_Q Err}^2_{\kappa,\Theta_2(a')}(x, \sum_i r_i^2))).\end{aligned} \tag{7.49}$$

If $\lambda' = \infty$, this holds trivially for large enough a'. Let us then discuss the $\lambda' \neq \infty$ case. In what follows,

$$U(a', x) = \overline{\forall_Q Err}^2_{\kappa,\Theta_1(a')}(x, \sum_i r_i^2)) - \overline{\forall_Q Err}^2_{\kappa,\Theta_2(a')}(x, \sum_i r_i^2)) \tag{7.50}$$

and

$$V(a', x) = \overline{\forall_Q Err}^2_{\kappa,\lambda'}(x, \sum_i r_i^2)) - \overline{\forall_Q Err}^2_{\kappa,\Theta_2(a')}(x, \sum_i r_i^2)), \tag{7.51}$$

which gives

$$\begin{aligned}&U(a', x) - V(a', x) = \\ &\overline{\forall_Q Err}^2_{\kappa,\Theta_1(a')}(x, \sum_i r_i^2)) - \overline{\forall_Q Err}^2_{\kappa,\lambda'}(x, \sum_i r_i^2)) \geq 0\end{aligned} \tag{7.52}$$

(since we only need to consider the cases in which the mean square error of $\Theta_1(a')$ is greater than or equal to that of λ'). In section 7.5.2.1 it was demonstrated that the difference between the mean square errors of two methods $\lambda_1, \lambda_2 < \infty$ converges proportionally to $\frac{1}{x^2}$, which means that the difference (7.52) also converges at this rate. It follows that

$$U(a', x) \approx V(a', x) + \frac{c}{x^2} \Leftrightarrow \frac{1}{a'}U(a', x) \approx \frac{1}{a'}V(a', x) + \frac{c}{a'x^2}. \tag{7.53}$$

for some constant c and all large enough x. Let us then examine the conditions under which

$$V(a', x) > \frac{1}{a'}V(a', x) + \frac{c}{a'x^2}, \tag{7.54}$$

holds. Since $V(a', x) \geq 0$, the inequality (7.54) entails

$$a' > 1 + \frac{c}{x^2 V(a', x)} \tag{7.55}$$

while the difference $V(a', x)$ converges proportionally to $\frac{1}{x^2}$, which means that (7.55) equals some constant number. Hence, for some a' satisfying (7.55) and all large enough x, it follows that

$$\frac{1}{a'}U(a', x) < V(a', x). \tag{7.56}$$

Hence, the inequality (7.49) holds for some a' satisfying (7.55) and all large enough x. Note that this result is true no matter the magnitude of the difference $\Theta_1(a') - \Theta_2(a')$.

However, in each inequality above, the samples already obtained play a role. Any sample $w_{a'}$ has the effect that the relevant performance criterion is not the mean square error for sample sizes $x > a'$. Instead, one has to calculate the mean square error for samples which are extensions of $w_{a'}$.

This means that, for example, the errors of λ' with the two different samples are not reduced away from the inequality corresponding to (7.48) and an equation corresponding to (7.49) is thus not obtained. However, the additional term of the below form denoting the difference between the relevant errors of λ' for the samples yielding $\Theta_1(a')$ and $\Theta_2(a')$ as multiplied by $\frac{1}{x}$ is added *on the left side* of the equation corresponding to (7.49):

$$\frac{1}{a'}(Error(a', w^1_{a'}, \lambda') - Error(a', w^2_{a'}, \lambda')). \tag{7.57}$$

Since the degree of order of the latter sample (i.e., $w^2_{a'}$ corresponding to $\Theta_2(a')$) is assumed to be closer to $\sum_i r_i^2$ (recall that the cases in which $\Theta_1(a')$ closer to the optimum method than $\Theta_2(a')$ can be left out from our considerations), the sample $w^2_{a'}$ probably resembles the universe more than $w^1_{a'}$, at least for large a'. Hence, it can be assumed that the additional term of the form (7.57) is positive and one can thus adhere to the equation (7.49) modified by the effects of the two samples.

Recall the formula (7.10) for the error of the estimate (7.1) with respect to Q_i. It follows from this that the average square error of λ with respect to all Q-predicates is

$$\forall_Q Err^2_{\lambda,\kappa}(x, X_1, ..., X_\kappa, r_1, ..., r_\kappa) = \frac{1}{\kappa}\sum_{i=1}^{\kappa}(\frac{X_i + \frac{1}{\kappa}\lambda}{x + \lambda} - r_i)^2. \tag{7.58}$$

Let A_i denote the absolute frequency of Q_i in $w_{a'}$ and let Y_i be a random variable for the corresponding frequency in the part of the sample not covered by $w_{a'}$. The square error (7.58) takes the following form:

$$\frac{1}{\kappa}\sum_{i=1}^{\kappa}(\frac{A_i + Y_i + \frac{1}{\kappa}\lambda}{x + \lambda} - r_i)^2. \tag{7.59}$$

The expression (7.59) equals

$$\frac{1}{\kappa}\sum_{i=1}^{\kappa}(\frac{A_i}{x+\lambda}+\frac{Y_i+\frac{1}{\kappa}\lambda}{x+\lambda}-r_i)^2= \\ \frac{1}{\kappa}\sum_{i=1}^{\kappa}[(\frac{A_i}{x+\lambda})^2+2\frac{A_i}{x+\lambda}(\frac{Y_i+\frac{1}{\kappa}\lambda}{x+\lambda}-r_i)+(\frac{Y_i+\frac{1}{\kappa}\lambda}{x+\lambda}-r_i)^2]. \tag{7.60}$$

Consider the first term in the expression under the summation sign. The difference between terms of this form for $\lambda_1 < \infty$ and $\lambda_2 < \infty$ is

$$(\frac{A_i^1}{x+\lambda_1})^2-(\frac{A_i^2}{x+\lambda_2})^2 \tag{7.61}$$

for two different samples. It is clear that this difference converges at a rate proportional to $\frac{1}{x^2}$. Consider then the second term under the summation sign. Since the maximum absolute value of

$$\frac{Y_i+\frac{1}{\kappa}\lambda}{x+\lambda}-r_i \tag{7.62}$$

is 1, the convergence rate of the difference between terms of the second form for $\lambda_1 < \infty$ and $\lambda_2 < \infty$ for large x depends on

$$\frac{A_i^1}{x+\lambda_1}-\frac{A_i^2}{x+\lambda_2}=\frac{A_i^1x+A_i^1\lambda_2-A_i^2x-A_i^2\lambda_1}{(x+\lambda_1)(x+\lambda_2)}, \tag{7.63}$$

which converges in proportion to $\frac{1}{x}$. This is also the convergence rate of one term of the second form. Since the mean square error for $\lambda < \infty$ converges in proportion to $\frac{1}{x}$ although the square error does not in general converge to zero, it seems credible that a term of the second form in (7.59) converging proportionally to $\frac{1}{x}$ does not affect the convergence rate of the difference between the means of two errors of the form (7.59).

The remaining term, in turn, gives the square error for the part of the sample which is not included in A_i, $i = 1, ..., \kappa$. It follows that the difference between the expected values of errors of the form (7.59) for $\lambda_1 < \infty$ and $\lambda_2 < \infty$ decreases in proportion to $\frac{1}{x^2}$. Since the results without assuming the sample $w_{a'}$ earlier in this section also obtain no matter the magnitude of the difference between the errors at a', they can readily be generalized to the case with the sample $w_{a'}$.

The $\Theta(\lambda_a, a', w_{a'}) = \infty$ case, i.e., $d_o(w_{a'}) = \frac{1}{\kappa}$ is handled by simply declaring that should this be the case, the method to be used being that which is closest to $\lambda = \infty$ while being closer to λ' than $\lambda = \infty$. This is an *ad hoc* solution, but sufficient for the present discussion.

However, it has not yet been demonstrated that Θ performs better than λ'. The question of finding the values a' and large enough x referred to above will be discussed in section 7.5.3.4 below. If these values can be effectively found, one can use the given constant method λ' until some x and then switch to $\Theta(\lambda_a, a', w_{a'})$ at x. The mean square error of this combination of inductive methods would then be smaller than that of λ'.

The other question relates to the original idea behind Θ. Stopping the updating process at $\Theta(\lambda_a, a', w_{a'})$ does not fully match this idea. However, due to technical complexities the question of whether the continuous updating of the inductive method would perform better than λ' will not be addressed in this study.

7.5.3.4 Finding the sample size

One can choose $\lambda_a = \lambda'$, use this λ_a up to some x, and at this x switch to $\Theta(\lambda_a, a', w_{a'})$. This combination of rules would correspond to the generalized form of Θ mentioned in 1) on p. 110. Obviously such a rule – in the effective sense of a computable rule – does not exist unless one provides an effective method for computing the values of a' and x. However, for obtaining a rule like this it suffices to produce *some* values for a' and x, which are not necessarily their least possible values.

The difficulty here is that the difference between λ' and the optimum method is not known for a given numerical value of λ' unless $\sum_i r_i^2$ is given. On the other hand, the formulation of the generalized Θ-rule obviously cannot presuppose knowledge of $\sum_i r_i^2$; if $\sum_i r_i^2$ was knowable, the whole problem of choosing the optimum inductive method would vanish. Nevertheless, the sample size up to which one must adhere to λ' cannot be calculated without $\sum_i r_i^2$. This means that one has to *assume* that d_1 in (7.39) differs from $\sum_i r_i^2$ at least with some $Y > 0$.

Under these circumstances, it seems possible that a rule for computing the values of a' and x fulfilling the above requirements can be formulated such that it does not contain $\sum_i r_i^2$ as a parameter, although some additional work is required to obtain its exact formulation.

However, it is possible that in reality $|d_1 - \sum_i r_i^2| < Y$ holds, i.e., d_1 is closer to the degree of order of the universe than Y. The procedure of calculating a' and x on the basis of Y then does not yield a correct result and the expected performance of $\Theta(\lambda_a, a', w_{a'})$ is not better than that of λ' for samples greater or equal to x.

It is thus clear that given an arbitrary λ', the generalized Θ-rule does not demonstrably perform better than λ'. Nevertheless, it can be argued that it performs better than any given λ' for which

$$|\delta(\lambda') - \sum_i r_i^2| \geq Y, \tag{7.64}$$

holds, where Y is an arbitrarily small non-zero natural number. In other words, if the chosen $\lambda_a = \lambda'$ fulfils the above condition, the generalized Θ-rule will perform better than λ'.

Observe that the generalized Θ-rule does not necessarily perform better than every method λ' for which (7.64) holds. The result argued for here can hold only for a λ' for which $\lambda_a = \lambda'$; there may certainly also be other methods for which (7.64) holds and which are, in addition, closer to the optimum method than λ'. Therefore, it is only argued here that, when one deliberates about whether to use a particular λ', the generalized Θ-rule performs better than this λ' and should thus be preferred to it.

To sum up the above discussion: given a numerical value of λ', the generalized Θ-rule is not argued to perform better always. On the other hand, the rule (when it is fully formalized) does perform better than a method which has at least a given distance from the optimum method. This distance can be chosen to be arbitrarily small. Hence, it can be argued that the general form of the Θ-rule performs better than any given inductive method except an arbitrarily small set of them, namely, the set of those for which (7.64) holds. Which methods actually belong to this set is not known unless the degree of order of the universe is known.

7.5.3.5 The inverse inference

The foregoing discussion presented an argument in favour of the generalized Θ-rule, a modified version of the Θ-rule. This section continues the discussion on the justification of the Θ-rule, analysing the relation between the Θ-rule and the statistical inference from a sample to the population. It is held that the justification of the Θ-rule does not require that one should be justified in drawing conclusions from a given sample to the whole population.

A well-known statistical formulation of the problem of induction is the following: given a random sample in which the frequency of the predicate Q is r_x, what is the probability that Q has a frequency which is within some given interval from r_x in the whole population? This is often called inverse inference in the philosophical literature.[9]

In direct inference, in contrast to inverse inference, one draws conclusions about a random sample based on knowledge about the whole population. The argumentation in the previous section resembles direct inference in the sense that one discusses probabilities of obtaining various kinds of samples from the universe (although assuming any particular kind of universe is irrelevant for the argument). For example, the determination of the mean square error of a particular

[9]Various forms of inductive inference, including inverse inference, are listed in Carnap 1962, pp. 207-208.

inductive method in a given universe and sample size is based on using direct inference.

Although one can raise doubts about whether a sample is ever random in real life or whether a sample can be known to be a random one, direct inference is usually considered to be less problematic than inverse inference; in direct inference, one does not attribute properties to the whole population on the basis of observations concerning only a part of it, whereas this is exactly what is done in inverse inference.

Let us now consider how inverse inference proceeds in practice. Suppose the relative frequency of Q, r_x, is close to r in the sample w_x of size x. If x is large, it would seem intuitively justified to infer that the relative frequency of Q must also be close to r_x in the whole population w_u, if not necessarily, at least with high probability. However, it is well known that this conclusion is not valid without further qualifications, as can be seen from the following illustration using Bayes's formula:

$$\begin{aligned} P(r_u \approx r | r_x \approx r) = \frac{P([r_u \approx r] \& [r_x \approx r])}{P(r_x \approx r)} = \\ \frac{P([r_u \approx r] \& [r_x \approx r])}{P(r_u \approx r)} \cdot \frac{P(r_u \approx r)}{P(r_x \approx r)} = \\ \frac{P(r_x \approx r | r_u \approx r) \cdot P(r_u \approx r)}{P(r_x \approx r)}. \end{aligned} \tag{7.65}$$

Here r_u is the relative frequency of Q in the whole universe w_u and

$$P(r_x \approx r | r_u \approx r) \tag{7.66}$$

corresponds roughly to the probability derived in Carnap (1952) (cf. section 7.3). If (7.66) is given, one can make the inverse inference to the population illustrated by (7.65), but only assuming that the prior probabilities $P(r_u \approx r)$ and $P(r_x \approx r)$ are known – and in general they are not. However, the relation between the Θ-rule and the inverse inference cannot be discussed in any further detail here.

7.5.3.6 The convergence rate of the variance

What remains of the discussion is the proof that $Var(d_o(w_x))$ converges toward zero when $x \to \infty$ at least at a speed of convergence proportional to $\frac{1}{x}$.

Observe first that

$$Var(d_o(w_x)) = E([d_o(w_x)]^2) - [E(d_o(w_x))]^2. \tag{7.67}$$

This equals

$$E([\sum_i (\frac{X_i}{x})^2]^2) - [E(\sum_i (\frac{X_i}{x})^2)]^2. \tag{7.68}$$

Consider the first term above. One obtains

$$E([\sum_i (\frac{X_i}{x})^2]^2) = \frac{1}{x^4} E(\sum_i X_i^2 \sum_i X_i^2) = \tag{7.69}$$
$$\frac{1}{x^4} E(\sum_i \sum_j X_i^2 X_j^2) = \frac{1}{x^4} \sum_i \sum_j E(X_i^2 X_j^2).$$

Consider then the second term in (7.68):

$$[E(\sum_i (\frac{X_i}{x})^2)]^2 = [\frac{1}{x^2} \sum_i E(X_i^2)]^2 = \tag{7.70}$$
$$\frac{1}{x^4} \sum_i E(X_i^2) \sum_i E(X_i^2) = \frac{1}{x^4} \sum_i \sum_j E(X_i^2) E(X_j^2). \tag{7.71}$$

Substitution in (7.68) yields

$$Var(d_o(w_x)) = \frac{1}{x^4} [\sum_i \sum_j E(X_i^2 X_j^2) - \sum_i \sum_j E(X_i^2) E(X_j^2)] \tag{7.72}$$
$$= \frac{1}{x^4} \sum_i \sum_j [E(X_i^2 X_j^2) - E(X_i^2) E(X_j^2)]. \tag{7.73}$$

Schwartz's inequality says that

$$|E(XY)| \leq \sqrt{E(X^2) E(Y^2)}. \tag{7.74}$$

Hence,

$$|E(X_i^2 X_j^2)| \leq \sqrt{E(X_i^4) E(X_j^4)} \tag{7.75}$$
$$\Leftrightarrow [E(X_i^2 X_j^2)]^2 \leq E(X_i^4) E(X_j^4).$$

It follows that

$$E(X_i^2 X_j^2) - E(X_i^2) E(X_j^2) \leq \sqrt{E(X_i^4) E(X_j^4)} - E(X_i^2) E(X_j^2) \tag{7.76}$$

Now assume that the convergence of (7.72) to zero when $x \to \infty$ can be proved when $E(X_i^2 X_j^2)$ is replaced by $\sqrt{E(X_i^4) E(X_j^4)}$. This means that even an expression with possibly greater individual term values than in (7.72) tends to zero.

Note also that since (7.72) is a variance, it is always zero or positive. Hence, it also follows that (7.72) tends to zero.

The factor $\frac{1}{x^4}$ can be inserted into the expression (7.72). Let us now show that

$$\frac{1}{x^4}[\sqrt{E(X_i^4)E(X_j^4)} - E(X_i^2)E(X_j^2)] = \frac{1}{x^4}\sqrt{E(X_i^4)E(X_j^4)} - \frac{1}{x^4}E(X_i^2)E(X_j^2) \tag{7.77}$$

tends to zero when $x \rightarrow \infty$ at a rate of convergence at least proportional to $\frac{1}{x}$. It will be shown that this holds if

$$\frac{1}{x^8}E(X_i^4)E(X_j^4) - \frac{1}{x^8}[E(X_i^2)]^2[E(X_j^2)]^2 = \frac{1}{x^4}E(X_i^4)\frac{1}{x^4}E(X_j^4) - \frac{1}{x^4}[E(X_i^2)]^2\frac{1}{x^4}[E(X_j^2)]^2 \tag{7.78}$$

tends to zero at this speed of convergence. Let now $a = \sqrt{E(X_i^4)E(X_j^4)} \geq 0$, $b = E(X_i^2)E(X_j^2) \geq 0$ and $c = \frac{1}{x^4}$. Then consider the convergence of $c^2(a^2 - b^2) = c^2(a - b)(a + b) = (ca - cb)c(a + b)$. The factor $c(a + b)$ equals

$$\frac{1}{x^4}[\sqrt{E(X_i^4)E(X_j^4)} + E(X_i^2)E(X_j^2)]. \tag{7.79}$$

This expression does not tend to zero unless $E(X_i^4)$ or $E(X_j^4)$ is zero and $E(X_i^2)$ or $E(X_j^2)$ is zero, in which case it is trivial that (7.77) converges to zero at least at the rate of $\frac{1}{x}$. If this is not the case, the expected values in (7.79) must be determined by the proportions of the corresponding Q-predicates in the universe, which clearly entails that (7.79) does not converge toward zero. Hence, $(ca - cb)$ must converge to zero at least at the rate of $\frac{1}{x}$ and the desired result obtains.

Moreover, observe that for arbitrary a, b, c, d it holds that

$$ab - cd = ab - cb + bc - cd = (a - c)b + (b - d)c. \tag{7.80}$$

Hence, when h_1, h_2 are constants, if $(a - c) \rightarrow 0$ and $(b - d) \rightarrow 0$ at a rate of convergence proportional to $\frac{1}{x}$, and $b \rightarrow h_1, c \rightarrow h_2$, then $(ab - cd) \rightarrow 0$ at a speed of convergence proportional to $\frac{1}{x}$. Let now $a = \frac{1}{x^4}E(X_i^4)$, $b = \frac{1}{x^4}E(X_j^4)$, $c = \frac{1}{x^4}[E(X_i^2)]^2$ and $d = \frac{1}{x^4}[E(X_j^2)]^2$. If, when $x \rightarrow \infty$,

$$\frac{1}{x^4}E(X_i^4) - \frac{1}{x^4}[E(X_i^2)]^2 \rightarrow 0 \tag{7.81}$$

and

$$\frac{1}{x^4}E(X_j^4) - \frac{1}{x^4}[E(X_j^2)]^2 \rightarrow 0 \tag{7.82}$$

at a speed of convergence proportional to $\frac{1}{x}$, plus if $\frac{1}{x^4}E(X_j^4)$ and $\frac{1}{x^4}[E(X_i^2)]^2$ approach some constant numbers, then (7.78) tends to zero at a speed of convergence proportional to $\frac{1}{x}$, which means that (7.72) tends to zero at least a speed of convergence proportional to $\frac{1}{x}$.

It thus has to be shown that (7.81) holds at a speed of convergence proportional to $\frac{1}{x}$ for an arbitrary i and that $\frac{1}{x^4}E(X_j^4)$ and $\frac{1}{x^4}[E(X_i^2)]^2$ approach some constant numbers. Once this has been proved, it has clearly been shown that (7.67) tends to zero at the speed of convergence proportional to $\frac{1}{x}$.

Consider first the term $E(X_i^4)$.

One can express X_i^2 by using indicator variables (1_y is used as short for $1_i(y)$):

$$(\sum_{y=1}^{x} 1_y)^2 = \sum_{y=1}^{x} 1_y^2 + \sum_{y\neq z}^{x} 1_y 1_z = \sum_{y=1}^{x} 1_y + \sum_{y\neq z}^{x} 1_{yz}, \tag{7.83}$$

where $1_y 1_z = 1_{yz}$ and $\sum_{y\neq z}^{x} 1_{yz}$ is the sum of all 1_{yz} where $y \neq z$. It follows that

$$X_i^3 = (\sum_{y=1}^{x} 1_y)^3 = \sum_{y=1}^{x} 1_y (\sum_{y=1}^{x} 1_y + \sum_{y\neq z}^{x} 1_{yz}) \tag{7.84}$$
$$= \sum_{y=1}^{x} 1_y + \sum_{y\neq z}^{x} 1_{yz} + \sum_{v=1}^{x} 1_v \sum_{y\neq z}^{x} 1_{yz}.$$

Thus

$$X_i^4 = (\sum_{y=1}^{x} 1_y)^4 = \sum_{y=1}^{x} 1_y (\sum_{y=1}^{x} 1_y + \sum_{y\neq z}^{x} 1_{yz} + \sum_{v=1}^{x} 1_v \sum_{y\neq z}^{x} 1_{yz}) \tag{7.85}$$
$$= X_i^2 + \sum_{v=1}^{x} 1_v \sum_{y\neq z}^{x} 1_{yz} + (\sum_{u=1}^{x} 1_u)^2 \sum_{y\neq z}^{x} 1_{yz}.$$

Consider the last term of the last form above,

$$(\sum_{u=1}^{x} 1_u)^2 \sum_{y\neq z}^{x} 1_{yz}. \tag{7.86}$$

One of its factors is

$$\sum_{u=1}^{x} 1_u \sum_{y\neq z}^{x} 1_{yz} = (1_1 + ... + 1_x) \sum_{y\neq z}^{x} 1_{yz}, \tag{7.87}$$

which also equals the middle term in the last form of (7.85). Each term 1_h in $1_1 + ... + 1_x$ has a different effect on the terms in $\sum_{y\neq z}^{x} 1_{yz}$, depending on whether

h equals one of y or z or not. If $h = y$ or $h = z$, 1_h is redundant in the term $1_h 1_y 1_z$. Consider now the terms in (7.87) for which this holds. For $h = y$ or $h = z$ in (7.87), the result of multiplying $\sum_{y\neq z}^{x} 1_{yz}$ by 1_h is

$$\sum_{z\neq h}^{x} 1_{hz} + \sum_{y\neq h}^{x} 1_{yh} = 2\sum_{y\neq z}^{x} 1_{yz}. \tag{7.88}$$

The remaining terms of (7.87) are of the form $\sum_{v\neq y\neq z}^{x} 1_{vyz}$; hence, (7.87) equals

$$2\sum_{y\neq z}^{x} 1_{yz} + \sum_{v\neq y\neq z}^{x} 1_{vyz}. \tag{7.89}$$

The multiplication of (7.87) by the remaining factor of (7.86), $\sum_{u=1}^{x} 1_u$, gives

$$\begin{aligned}&\sum_{u=1}^{x} 1_u(2\sum_{y\neq z}^{x} 1_{yz} + \sum_{v\neq y\neq z}^{x} 1_{vyz}) \\ &= 4\sum_{y\neq z}^{x} 1_{yz} + 2\sum_{v\neq y\neq z}^{x} 1_{vyz} + \sum_{u=1}^{x} 1_u \sum_{v\neq y\neq z}^{x} 1_{vyz}.\end{aligned} \tag{7.90}$$

The procedure for computing the last term in (7.90) is familiar (see above), each 1_{vyz} being counted in three times: first with 1_v, then with 1_y and finally with 1_z. The result is obtained when this is added to terms where $u \neq v \neq y \neq z$. It follows that (7.90) equals

$$4\sum_{y\neq z}^{x} 1_{yz} + 2\sum_{v\neq y\neq z}^{x} 1_{vyz} + 3\sum_{v\neq y\neq z}^{x} 1_{vyz} + \sum_{u\neq v\neq y\neq z}^{x} 1_{uvyz}. \tag{7.91}$$

When put together, the above yields

$$\begin{aligned}&X_i^4 = X_i^2 + 2\sum_{y\neq z}^{x} 1_{yz} + \sum_{v\neq y\neq z}^{x} 1_{vyz} + 4\sum_{y\neq z}^{x} 1_{yz} + 2\sum_{v\neq y\neq z}^{x} 1_{vyz} + \\ &3\sum_{v\neq y\neq z}^{x} 1_{vyz} + \sum_{u\neq v\neq y\neq z}^{x} 1_{uvyz} = \\ &X_i^2 + 6\sum_{y\neq z}^{x} 1_{yz} + 6\sum_{v\neq y\neq z}^{x} 1_{vyz} + \sum_{u\neq v\neq y\neq z}^{x} 1_{uvyz}.\end{aligned} \tag{7.92}$$

Let us now proceed to calculate $\frac{1}{x^4}E(X_i^4)$, which was set as the objective on p. 128. Observe first that $E(1_y) = P\{1_y = 1\} = r_i$ (recall the abbreviation $1_y = 1_i(y)$). Since 1_y and 1_z are independent if $y \neq z$,

$$E(1_{yz}) = E(1_y 1_z) = E(1_y)E(1_z) = r_i^2 \tag{7.93}$$

and analogically for $E(1_{vyz})$ and $E(1_{uvyz})$.

The number of sequences of length k from the population of x elements is

$$\frac{x!}{(x-k)!}. \tag{7.94}$$

Hence, one can form

$$\frac{x!}{(x-j)!} \tag{7.95}$$

products of the form $1_1 ... 1_j$ when $j \leq x$. One thus obtains (using the above note about calculating expected values of indicator variables)

$$\begin{aligned}
&\frac{1}{x^4}E(X_i^4) = \frac{1}{x^4}(E(X_i^2) + 6\frac{x!}{(x-2)!}r_i^2 + 6\frac{x!}{(x-3)!}r_i^3 + \frac{x!}{(x-4)!}r_i^4) = \\
&\frac{1}{x^4}E(X_i^2) + 6(\frac{1}{x^2} - \frac{1}{x^3})r_i^2 + 6(\frac{1}{x} - \frac{3}{x^2} + \frac{2}{x^3})r_i^3 + \\
&(1 - \frac{6}{x^3} + \frac{11}{x^2} - \frac{6}{x})r_i^4.
\end{aligned} \tag{7.96}$$

It is clear from this that (7.96) converges toward a constant number.

Now one can proceed to calculate (7.81). By (7.36),

$$\begin{aligned}
\frac{1}{x^4}(E(X_i^2))^2 &= \frac{1}{x^4}(xr_i(1-r_i) + r_i^2x^2)^2 \\
&= r_i^4 + \frac{1}{x^2}r_i^2(1-r_i)^2 + \frac{2}{x}r_i^3(1-r_i).
\end{aligned} \tag{7.97}$$

When $x \to \infty$, (7.97) tends to r_i^4, which is a constant number. Hence, the conditions concerning $\frac{1}{x^4}(E(X_i^2))^2$ and $\frac{1}{x^4}E(X_i^4)$ on p. 128 are satisfied.

Let us proceed to show that (7.81) holds.

Because $\frac{1}{x^4}(E(X_i^2)) = \frac{1}{x^4}(xr_i(1-r_i) + r_i^2x^2)$, (7.96) reduces to

$$\begin{aligned}
&\frac{1}{x^4}(xr_i(1-r_i) + r_i^2x^2) + 6(\frac{1}{x^2} - \frac{1}{x^3})r_i^2 + 6(\frac{1}{x} - \frac{3}{x^2} + \frac{2}{x^3})r_i^3 + \\
&\qquad (1 - \frac{6}{x^3} + \frac{11}{x^2} - \frac{6}{x})r_i^4 = \\
&\frac{1}{x^3}[r_i - r_i^2 - 6r_i^3 + 12r_i^3 - 6r_i^4] + \frac{1}{x^2}[r_i^2 + 6r_i^2 + 18r_i^3 + 11r_i^4] + \\
&\qquad \frac{1}{x}[6r_i^3 - 6r_i^4] + r_i^4.
\end{aligned} \tag{7.98}$$

Recall that what was to be proved was the convergence toward zero of $\frac{1}{x^4}E(X_i^4) - \frac{1}{x^4}E(X_i^2)^2$ at a speed of convergence proportional to $\frac{1}{x}$. Observe that the terms r_i^4 in (7.97) and (7.98) cancel each other out. The rest of the terms tend to zero at least the rate of $\frac{1}{x}$ in both expressions when $x \to \infty$, from which the desired result follows. QED.

7.5.3.7 Example: the probability of uniform evidence

Employing the Θ-rule leads to an interesting result concerning the prior probability of an infinite and uniform stream of evidence data. Observe, however, that the constructive validity of the methods employed in this section is not discussed.

For the extreme method $\lambda = 0$, the prior probability of a uniform stream of data is trivially 1. For the other extreme $\lambda = \infty$, the prior probability of a uniform stream of data must clearly tend to zero since no piece of evidence can make future evidence of the same kind more probable. Hence, this section discusses only the interesting cases $0 < \lambda < \infty$.

The optimum λ-method for uniform evidence (i.e., evidence representing a single Q-predicate) is the straight rule $\lambda = 0$. After obtaining uniform evidence, even consisting of a single individual, the output of the Θ-rule (7.23) is precisely the straight rule. However, it is usually thought that prior considerations about the correct inductive method have some weight in choosing the method. In most cases, it is absurd to assign a probability of 1 to the next individual being similar as the first observed one.

For this reason, a more general form of the Θ-rule (7.23) will now be used.

The general rule is defined as follows:

$$\begin{cases} \Theta_{gen}(\lambda_a, c, 0, \_) = \lambda_a \\ \Theta_{gen}(\lambda_a, c, x+1, w_{x+1})) = & \Theta(\lambda_a, c, x, w_x) + \\ & \frac{1}{c}[\delta^{-1}(d_o(w_{x+1})) - \Theta(\lambda_a, c, x, w_x)] \end{cases} \tag{7.99}$$

where w_0 equals the empty sequence _.

Observe that for the considerations of this section it does not matter if the application of the Θ-rule starts only after a certain sample size (cf. the discussion in the previous chapter).

Method λ_a is the initial method based on prior considerations and c is a caution factor which indicates how much the degree of order should be adjusted at each step. The larger c is, the more cautious the adjustment is. For reasons of simplicity, it will be assumed that one has already moved the method away from the extreme method $c^\dagger$. Because the λ-value of $c^\dagger$ is ∞, no correction factor can adjust it. This means that the correction rule would have to be formulated for degrees of order,

which would result in more complicated calculations. Similarly, it is assumed that $d_o(w_{x+1}) \neq \frac{1}{\kappa}$.

Consider now the prior probability of consequently obtaining only heads when tossing a coin. Suppose the situation is modelled in a monadic language with one predicate, denoting the results of consecutive tosses of a coin. Moreover, suppose that the evidence reflects absolute uniformity, i.e., the tosses have been either all heads or all tails. The optimum inductive method in this case is the extreme method $\lambda = 0$, which means that the first application of the correction rule yields the method

$$\lambda_a(1 - \frac{1}{c}). \tag{7.100}$$

If the incoming evidence remains homogeneous, the resulting methods can be recursively calculated from the definition of (7.99). The result for a sample size k is given by

$$\lambda_a(1 - \frac{1}{c})^k. \tag{7.101}$$

The probability of an unlimited number of heads (or tails) is given by substituting the above in formula 11-4 given in Carnap (1952, p. 33) and forming the product over the indices $k \geq 1$:

$$\prod_{k=1}^{\infty} \frac{k + \frac{\lambda_a(1-\frac{1}{c})^k}{2}}{k + \lambda_a(1 - \frac{1}{c})^k}. \tag{7.102}$$

This formula can be compared to that for a fixed value of $0 < \lambda < \infty$:

$$\prod_{k=1}^{\infty} \frac{k + \frac{\lambda}{2}}{k + \lambda}. \tag{7.103}$$

It will be shown below that a routine convergence test, the ratio test, yields the result that (7.102) converges to a non-zero value but remains indifferent concerning the convergence of (7.103).

The terms of (7.103) can be written in the following form:

$$1 + \frac{1}{2}(\frac{k}{\lambda + k} - 1). \tag{7.104}$$

Provided that $a_k > 0$ or $a_k < 0$ from some value of k onwards, a product of the form

$$\prod_{k=1}^{\infty}(1 + a_k) \tag{7.105}$$

converges to a non-zero value if and only if the series

$$\sum_{k=1}^{\infty}(a_k) \tag{7.106}$$

converges. Since it was assumed that $\lambda > 0$, it holds that $(\frac{k}{\lambda+k} - 1) < 0$. Hence, as we see when the terms are written in the form (7.104), the product (7.103) converges to a non-zero value if and only if

$$\sum_{k=1}^{\infty}\frac{1}{2}(\frac{k}{\lambda+k} - 1) \tag{7.107}$$

converges, i.e., if

$$\sum_{k=1}^{\infty}(\frac{k}{\lambda+k} - 1) \tag{7.108}$$

converges.

Here one can apply the ratio test. If

$$\lim_{k\to\infty}|\frac{a_{k+1}}{a_k}|, \tag{7.109}$$

is smaller than 1, the series (7.106) converges, if greater than 1, the series diverges, and if (7.109) is equal to 1, the convergence remains undecided.

For (7.108), the ratio in (7.109) is equal to

$$\frac{\frac{k+1}{\lambda+k+1} - 1}{\frac{k}{\lambda+k} - 1}, \tag{7.110}$$

which reduces to

$$\frac{\lambda+k}{\lambda+k+1} = \frac{\frac{\lambda}{k}+1}{\frac{\lambda}{k}+\frac{1}{k}+1}. \tag{7.111}$$

From the above form it is clear that, when $k \to \infty$ (and $\lambda < \infty$, as has been assumed), the value of the ratio approaches 1. Hence, the ratio test leaves the question of the convergence of the series undecided.

Let us now examine the convergence of the product (7.102). It is assumed that $\lambda_a > 0$ since otherwise the prior probability of uniform evidence becomes 1 with the correction rule as well and the whole issue becomes trivial. Observe that when $\lambda_a > 0$, then $\lambda_a(1 - \frac{1}{c})^k > 0$.

The question of the convergence is now decided by the ratio test since the convergence of terms of the product towards 1 is faster in (7.102) than in (7.103), as is apparent when the terms of (7.102) are written in the form

$$1 + \frac{k}{2[\lambda_a(1 - \frac{1}{c})^k + k]} - \frac{1}{2}. \tag{7.112}$$

Let $y = 1 - \frac{1}{c}$. The formula (7.110) corresponding to (7.102) reads

$$\frac{\frac{k+1}{\lambda_a y^{k+1}+k+1} - 1}{\frac{k}{\lambda_a y^k+k} - 1}, \tag{7.113}$$

which can be reduced to

$$\frac{\lambda_a y^{k+1} + ky}{\lambda_a y^{k+1} + k + 1}. \tag{7.114}$$

This reduces to

$$\frac{\frac{\lambda_a y^{k+1}}{k} + y}{\frac{\lambda_a y^{k+1}}{k} + \frac{1}{k} + 1}. \tag{7.115}$$

As $k \to \infty$ (and $\lambda_a < \infty$), (7.115) does not tend to 1 since the numerator tends to y and the denominator to 1, and thus the ratio tends to $y = 1 - \frac{1}{c}$, which is less than 1. Hence, the product (7.102) converges to a non-zero value, which shows that an infinite stream of evidence consisting of occurrences of only one Q-predicate has a non-zero probability. Chapter 8 will examine what this means from the point of view of prior probabilities of universal generalizations.

The finding readily extends to other monadic languages and bodies of evidence that contain occurrences of only one Q-predicate.

7.5.4 The cumulative mean square error

This section provides a short digression to the definition of the mean cumulative square error, which is an interesting notion in its own right, even with no direct relevance to the proofs above.

Recall the definition of error of the λ-estimate, (7.10). The idea of the cumulative error is to consider the sum error of consecutive estimates with nested samples.

The sum error for some values of X_i in two consecutive nested samples of the sizes x and $x + 1$ can be written as

$$Err_{\lambda,\kappa}(x, X_i, r_i) + Err_{\lambda,\kappa}(x, X_i + 1_{i(x+1)}, r_i). \tag{7.116}$$

The expected value of (7.116) is

$$E[Err_{\lambda,\kappa}(x, X_i, r_i) + Err_{\lambda,\kappa}(x, X_i + 1_{i(x+1)}, r_i)] =$$
$$E[Err_{\lambda,\kappa}(x, X_i, r_i)] + E[Err_{\lambda,\kappa}(x, X_i + 1_{i(x+1)}, r_i)].$$

Observe that the expected value operator can be applied to the two error expressions separately even though the errors are not independent of each other because the general formula $E(X + Y) = E(X) + E(Y)$ does not require that X and Y be independent random variables.

One can define the following formula for the *cumulative square error* for nested samples beginning from size of 1 up to k:

$$CErr^2_{\lambda,\kappa}(k, X_i, r_i) = \sum_{x=1}^{k} [Err_{\lambda,\kappa}(x, X_i, r_i)]^2. \tag{7.117}$$

The mean or expected value of the cumulative square error for the predicate Q_i is, by reference to (7.17),

$$E(CErr^2_{\lambda,\kappa}(k, 1_{i(1)}, ..., 1_{i(k)}, r_i)) = \sum_{x=1}^{k} \overline{Err}^2_{\lambda,\kappa}(x, r_i), \tag{7.118}$$

i.e., it equals the sum of mean square errors for the nested samples from $x = 1$ to $x = k$ for the predicate Q_i. One can thus use the terms mean cumulative square error and cumulative mean square error interchangeably.

With respect to all Q_i, the average mean cumulative square error (referred to in what follows only as the mean cumulative square error or cumulative mean square error) is

$$\frac{1}{\kappa} \sum_{i=1}^{\kappa} \sum_{x=1}^{k} \overline{Err}^2_{\lambda}(x, r_i) = \sum_{x=1}^{k} \frac{1}{\kappa} \sum_{i=1}^{\kappa} \overline{Err}^2_{\lambda}(x, r_i). \tag{7.119}$$

Using the form (7.20), (7.119) equals to

$$\sum_{x=1}^{k} \overline{\forall_Q Err}^2_{\kappa,\lambda}(x, \sum_i r_i^2), \tag{7.120}$$

which will also be denoted by

$$\overline{\forall_Q CErr}^2_{\kappa,\lambda}(x, \sum_{i=1}^{\kappa} r_i^2). \tag{7.121}$$

Observe that

$$\sum_{x=1}^{k} \frac{1}{\kappa} \sum_{i=1}^{\kappa} \overline{Err}_{\lambda}^{2}(x, r_i) = \sum_{x=1}^{k} \frac{1}{\kappa} \sum_{i=1}^{\kappa} E([Err_{\lambda}(x, X_i, r_i)]^2) = \tag{7.122}$$

$$E(\sum_{x=1}^{k} \frac{1}{\kappa} \sum_{i=1}^{\kappa} [Err_{\lambda}(x, X_i, r_i)]^2),$$

which means that one can also compose the expression for the cumulative square error with respect to all Q-predicates first and then apply the expected value operator to this expression.

Chapter 8

The correction rule and time

This section will discuss some implications of adopting the correction rule from a constructive point of view. It will turn out that, beside the effect on probabilities of pieces of empirical evidence, it also changes the way the prior probabilities of sentences are conceived of.

The first observation is that the correction rule θ does not seem quite general, since evidence does not necessarily consist of observations of single individuals. In inductive logic, any sentence of the language can function as evidence. This suggests that one needs a correction rule that would update the method with any kind of evidence, for example, ones that describe several individuals.

However, a simple argument shows that this would lead to difficulties. Consider a rule like θ above but with the variable x denoting not the number of individuals in the sample, but the number of evidence statements that have been observed. It is assumed that each of the evidence statements describes the properties of a finite number of individuals exhaustively.

It is easy to see that the results of applying the general form of the correction rule (7.99) with the evidence $H(i)\&H(i+1)$ at one go, and with two consecutive pieces of evidence $H(i)$ and $H(i+1)$ separately, would differ from each other (assuming that only the property H has previously occurred). In both of these ways of updating, the optimum method after obtaining the evidence remains $\lambda = 0$, but in the latter case, the current method is updated twice instead of only once, as in the first case.

This sensitivity to the time factor in obtaining evidence has important consequences. Consider the probability of uniform evidence discussed above. This probability was shown to be positive and non-infinitesimal, while the prior probability of a universal quantification with the same propositional content is zero.

The time factor will be discussed in more detail below.

8.1 Time and obtaining evidence

It seems that the optimum method depends on the duration of time in which the pieces of evidence are obtained, but is this a plausible view about probability considering that one arrives at the same body of knowledge about the world by obtaining e_1 first and then e_2 as by obtaining $e_1 \& e_2$ at one go? Should not these two updatings then result in equal posterior probabilities?

It has been argued above that it is rational to make adjustments to the inductive method in the course of the process of inquiry. For example, even if one starts with the method $\lambda = \infty$, which gives no weight to the empirical factor in determining the probabilities, very uniform evidence attributing a property Q to a large number of consecutively observed individuals should make a rational agent shift to another value of λ at some point of the inquiry. Since such an adjustment increases the probability of observing further individuals with the property Q, the probability of doing so is higher than it was before the adjustment. Hence, because a process of inquiry often involves a time factor, probabilities for a research process differ from ordinary conditional probabilities.

The status of evidence which is obtained piecewise is, in fact, different from that which is obtained at one go, in terms of knowledge in hand after obtaining the evidence. Even though the meaning of the evidence statements is the same, their status differs because they were obtained under different background conditions, determined by the different inductive methods. To accommodate these ideas into the framework of conditional probabilities, the evidence statement should inform us, not merely about the propositional content of the evidence, but whether it is known as well, and if so, whether it was obtained piecemeal or at one go. Thus, for example, the probability of an infinite body of evidence representing an infinite number of consecutive tosses resulting in heads, should be ascribed a different probability than the corresponding universal generalization that says that all the tosses will be heads.

However, it will be shown below that the above distinctions do not apply when truth is interpreted constructively, which means – more precisely – that universal generalization is assigned the same a priori probability as the corresponding stream of observational evidence.

8.1.1 Prior probability

Consider two sentences, $Q_1(1)$ and $Q_1(2)$, in which Q_i is an observable Q-predicate. Since constructive truth equals knowability, and knowability in this case is observability, $Q_1(1)$ and $Q_1(2)$ cannot be constructively true without being observable. The question of what it means for a sentence to be knowable or observable remains.

The actualist conception of constructive truth is one answer to the question. It will be shown below that the actualist conception entails that universal generalizations have the same probability as the corresponding body of evidence. However, the actualist conception may not be the most intuitive interpretation of constructive truth. One may wish to tackle the meaning of knowability in some other way. It will be shown in section 8.1.1.2 that a non-actualist conception of constructive truth also entails that universal generalizations and the corresponding streams of evidence have equal probabilities.

8.1.1.1 Actualist truth

One answer is to reject the question and to admit that constructive truth is a tensed conception, i.e., that being true means having been proved, which is often referred to as the actualist conception of constructive truth (for constructive conceptions of truth, see, e.g., Raatikainen 2004).

The following reasoning supports the actualist interpretation of constructive truth.

The statement that a proposition can be proved cannot, in the constructive setting, mean anything beyond there being a justification for the expression 'S can be proved'. A situation in which S is provable and it is not justified to say that S is provable cannot occur; in other words, S cannot be provable in some objective sense, without the justification for saying that S is provable. Holding that such a situation could occur would be a commitment to classical meaning theory concerning the expression 'S can be proved'.

If S has been proved, it seems intuitive that S was already provable before the proof was actually carried out. The above argument entails that it is not justified to say that S is provable before it is proved, but is it justified to say that, in a situation where S has actually been proved, it was provable before it was proved?

Saying that S was provable means that a situation obtained where the statement 'S is provable' was true. However, the actualist conception of constructive truth entails that S must be proved whenever 'S is provable' is true. Hence, S could not have been provable before it was proved, which is clearly an unintuitive feature in the actualist conception of constructive truth.

Let us now proceed to discuss what the actualist conception of constructive truth means from the point of view of probabilities.

The assertion that $Q_1(1)\&Q_1(2)$ is true, for instance, means that $Q_1(1)\&Q_1(2)$ is knowable, which in turn means that $Q_1(1)$ and $Q_1(2)$ are knowable. In the actualist conception of constructive truth, $Q_1(x)$ is observable only if it has been observed. Hence, $Q_1(1)\&Q_1(2)$ is true only if $Q_1(1)$ and $Q_1(2)$ have been observed.

Provided that an observation of a Q-property must last for a certain duration

of time, $Q_1(1)$ and $Q_1(2)$ can simultaneously have the status of having been observed, but they cannot have been observed at the same time. Hence, the (actualist) constructive meaning of the probability of $Q_1(1)\&Q_1(2)$ is the probability of a state of affairs in which they have been observed separately, one after the other.

But if one observes $Q_1(1)$ or any other Q-property, the conception about the degree of order of the universe is updated. It was argued above that it is rational to update the inductive method whenever new individuals with the same Q-property are observed. If the method is updated after observing $Q_1(1)$, the probability of $Q_1(2)$ changes from what it was before observing $Q_1(1)$. Hence, the probability of $Q_1(2)$ is different in the situation where no Q-individuals have been observed from its probability in the situation where $Q_1(1)$ has been observed.

Hence, because of the updating of the method after each observation of a Q-individual, universal generalizations are assigned the same prior probability as the corresponding stream of evidence.[1]

8.1.1.2 Non-actualist truth

In a non-actualist conception of constructive truth, a sentence is true if it is provable (in the sense of there being a method of proving it), without having to be actually proved.

In this interpretation of truth, S can be true timelessly, but not independently of knowability. Expressed in terms of observability, S is true if it is observable, which means that $Q_1(1)\&Q_1(2)$ is true iff $Q_1(1)$ is observable and $Q_1(2)$ is observable. But even if $Q_1(1)$ and $Q_1(2)$ are both observable, they cannot be observed at the same time. $Q_1(1)$ and $Q_1(2)$ can be timelessly true in the sense of observability, but only under the condition that one of them can be observed before the other one.

The probability of the truth of a universal generalization thus means the probability of its instances being observable one after another. This entails that the probability of a universal generalization with a Q-predicate as the sentential matrix is calculated by first calculating the probability of the first instance and then multiplying this probability by the probability of the second instance while taking into account the appropriate change in the inductive method, etc.

Since the inductive method can be updated after each observation of a Q-individual, universal generalizations (with Q-predicates) are assigned the same probability as the corresponding stream of evidence in the case of non-actualist conception of constructive truth as well.

[1] Observe that no application of the correction rule needs to take place if truth is interpreted classically since $Q_1(1)\&Q_1(2)$ can be timelessly true 'out there', without one of the conjuncts having to be established before the other.

8.1.1.3 The order of conjuncts

The equation

$$P(Q_1(1)\&Q_1(2)) = P(Q_1(2)) \cdot P(Q_1(1)|Q_1(2)) = P(Q_1(2)\&Q_1(1)) \quad (8.1)$$

reflects the interpretation of probabilities without the effect of the correction rule since the probability function P (i.e., the inductive method) remains unchanged after observing $Q_1(2)$, as seen from the middle form of the equation.

When the correction rule is applied, the order of observations becomes significant. Provided that the order of the conjuncts in the probability statement denotes the order of observation, it is not necessarily true that

$$P(Q_i(x)\&Q_j(x+1)) = P(Q_j(x+1)\&Q_i(x)) \quad (8.2)$$

since it may, for example, be the case that observing Q_i does not change the current method but observing Q_j does (in the latter case, $Q_i(x)$ is assigned a different probability than in the previous case).

As discussed in Ch. 7, there are grounds for choosing a correction rule instead of a constant method. Using a correction rule effects the probabilities of streams of evidential data. It can be argued (as has been done in this chapter) that in constructive semantics, these probabilities must be the same as the prior probabilities of the corresponding universal generalizations. However, the question of how the correction rule effects the findings concerning extendible truth and probability cannot be addressed in this study.

Bibliography

[1] Benenson, F.C. 1984: *Probability, Objectivity and Evidence.* Routledge & Kegan Paul.

[2] Billingsley, P. 1995: *Probability and Measure,* 3rd ed. Wiley, New York.

[3] Bishop, E. 1967: *Foundations of Constructive Analysis.* McGraw-Hill.

[4] Bricker, P. 1987: "Reducing possible worlds to language", *Philosophical Studies* 52, 331-355.

[5] Carnap, R. 1937: *The Logical Syntax of Language.* Routledge & Kegan Paul, London. Original German work: *Logische Syntax der Sprache.* Verlag von Julius Springer, Wien, 1934.

[6] Carnap, R. 1946: "Modalities and quantification". *Journal of Symbolic Logic* 11, 33-64.

[7] Carnap, R. 1947: *Meaning and Necessity.* The University of Chicago Press, Chicago.

[8] Carnap, R. 1952: *The Continuum of Inductive Methods.* The University of Chicago Press, Chicago.

[9] Carnap, R. 1962: *Logical Foundations of Probability,* 2nd edition (first edition in 1950). The University of Chicago Press, Chicago.

[10] Carnap, R. 1968: "Reply to J. Hintikka". In I. Lakatos (ed.), *The Problem of Inductive Logic.* North-Holland, Amsterdam, 312-314.

[11] Carnap, R. 1977: *Two Essays on Entropy.* Edited by A. Shimony. University of California Press.

[12] van Dalen, D. 1986: "Intuitionistic logic". In D. Gabbay & F. Guenthner (eds.), *Handbook of Philosophical Logic,* vol. III. D. Reidel, Dordrecht.

[13] De Finetti, B. 1972: *Probability, Induction and Statistics.* John Wiley & Sons.

[14] Douven, I. 2000: "Empirische toetsing van inductieve logica's", *Tijdschrift voor Filosofie* 62, 701-725.

[15] Dummett, M. 1977: *Elements of Intuitionism.* Oxford University Press, Oxford.

[16] Dummett, M. 1978: *Truth and Other Enigmas,* London, Duckworth.

[17] Fagin, R. 1976: "Probabilities on finite models". *The Journal of Symbolic Logic* 41, 50-58.

[18] Festa, R. 1994: *Optimum Inductive Methods. A Study in Inductive Probability, Bayesian Statistics, and Verisimilitude.* Kluwer Academic, Dordrecht.

[19] Field, H. 1977: "Logic, meaning, and conceptual role". *Journal of Philosophy* 74, 379-409.

[20] Fletcher, P. 2002: "A constructivist perspective on physics", *Philosophica Mathematica,* Series III, 10, 26-42.

[21] Grove A., Halpern J., & Koller D. 1996: "Asymptotic conditional probabilities: the non-unary case". *The Journal of Symbolic Logic* 61, 250-276.

[22] Harman, G. 1983: "Problems with probabilistic semantics". In A. Orenstein & R. Stern (eds.), *Developments In Semantics,* Haven Publications, 242-245.

[23] Heyting, A. 1934: *Matematische Grundlagenforschung. Intuitionismus. Beweistheorie.* Springer, Berlin.

[24] Hintikka, J. 1955: "Form and content in quantification theory". *Acta Philosophica Fennica* 8, 11-55.

[25] Hintikka, J. 1969: "Modality and Quantification". In J. Hintikka, *Models for Modalities,* D. Reidel, 57-70.

[26] Holm, R. 2003: "A constructive approach to state description semantics", *Journal of Applied Logic* 1, 13-46.

[27] Horwich, P. 1982: *Probability and Evidence.* Cambridge University Press, Cambridge.

[28] Jeffrey, R.C. 1971: "Probability measures and integrals". In R. Carnap & R.C. Jeffrey (eds.), *Studies in Inductive Logic and Probability.* University of California Press, 167-221.

[29] Keynes, J.M. 1921: *A Treatise on Probability.* Macmillan, London.

[30] Kolmogorov, A.N. 1933: *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Springer, Berlin.

[31] Kopylov, A. & Nogin, A. 2001: "Markov's principle for propositional type theory". In L. Fribourg (ed.), *Computer Science Logic, Proceedings of the 10th Annual Conference of the EACSL* , vol. 2142 of *Lecture Notes in Computer Science,* Springer-Verlag, 570-584.

[32] Kuipers, T. 1986: "Some estimates of the optimum inductive method". *Erkenntnis* 24, 37-46.

[33] Leblanc, H. 1983: "Alternatives to standard first-order semantics". In D. Gabbay & F. Guenthner (eds.), *Handbook of Philosophical Logic,* Vol. I., D. Reidel, 189-274.

[34] Lewis, D. 1971: "Immodest inductive methods". *Philosophy of Science* 38, 54-63.

[35] Lewis, D. 1974: "Spielman and Lewis on inductive immodesty". *Philosophy of Science* 41, 84-85.

[36] Liogon'kii, M.I. 1969: "On the conditional satisfiability ratio of logical formulas". *Mathematical Notes of the Academy of the USSR* 6, 856-861.

[37] Logue, J. 1991: "Weight of evidence, resiliency and second-order probabilities". In E. Eells & T. Maruszewski (eds.), *Probability and Rationality,* Rodopi, 147-172.

[38] Lynch, J. 1980: "Almost sure theories". *Annals of Mathematical Logic* 18, 91-135.

[39] Markov, A.A. 1962: "On constructive mathematics". *Trudy Matematicheskogo Instituta imeni V.A. Steklova* 67, 8-14. English Translation: *AMS Translations,* series 2, vol. 98, 1-9.

[40] Martin-Löf, P. 1968: *Notes on Constructive Mathematics.* Almqvist & Wiksell, Stockholm.

[41] Martin-Löf, P. 1990: “Mathematics of Infinity”. *Lecture Notes in Computer Science* 417, Springer-Verlag, 146-197.

[42] Pietarinen, J. 1974: "Inductive immodesty and lawlikeness". *Philosophy of Science* 41, 196-198.

[43] Raatikainen, P. 2004: "Conceptions of truth in intuitionism". *History and Philosophy of Logic* 25, 131-145.

[44] Ranta, A. 1992: "Worlds and state-descriptions", a talk given in the Finnish-Russian logic symposium.

[45] Skyrms, B. 1980: "Higher order degrees of belief". In D.H. Mellor (ed.), *Prospects for Pragmatism,*109-137.

[46] Spielman, S. 1972: "Lewis on immodest inductive methods". *Philosophy of Science* 39, 375-377.

[47] Troelstra, A. 1977: *Choice sequences. A Chapter of Intuitionistic Mathematics*. Oxford University Press, Oxford.

[48] Troelstra, A, & van Dalen, D. 1988: *Constructivism in mathematics. An introduction.* North-Holland, Amsterdam.

[49] Wright, C. 1992: *Truth and Objectivity.* Harvard University Press, Cambridge MA.