Approved by

First Reader
Judson Webb, Ph.D.
Associate Professor of Philosophy

Second Reader
Alisa Bokulich, Ph.D.
Associate Professor of Philosophy

Third Reader
William Rodi, Ph.D.
Principal Research Scientist, Geophysics, MIT
Acknowledgments

Prior to my enrollment in this degree program I had absolutely no previous education in philosophy. My learning curve, therefore, was steep and I had to cover a lot of ground in a short amount of time. It was not easy and it took away most of my social life for three years. Still, I consider myself enriched by the experience, and do not regret it. Therefore I want to thank all those who made this accomplishment possible for me. First, my advisor, Professor Judson Webb, who is a walking storehouse of philosophical information. I was always impressed by his knowledge of the foundations of mathematics and logic, two of my favorite subjects. Next, Professor Alisa Bokulich, who seemed to be able to refute any supposedly novel idea I thought I had by recalling a real historical counter-example. Finally I thank Bill Rodi, who has been on my committee for a graduate degree twice. His breadth of knowledge, both theoretical and applied, has been an inspiration. I can’t thank him enough for the many illuminating conversations we’ve had over the years.

I must also thank those who have no idea what I’ve been studying or why, but have supported me nonetheless. Most important among them are my father, William, my mother, Loana (watching over me from somewhere else...), my brother, David and his family, and all my close friends in Boston and elsewhere.

My greatest thanks, however, go to God, whom I don’t fully understand, but do fully love. I thank Him for this opportunity to return to school and ponder the deeper foundations of issues I find interesting. I pray that all the effort I put into writing this thesis, however flawed and imperfect it might be, will serve to glorify Him in some way.

Blessed is the man who meditates on wisdom
and who reasons intelligently.

Sirach 14:20
I argue that a probabilistic logical language incorporates all the features of deductive, inductive, and abductive inference with the exception of how to generate hypotheses ex nihilo. In the context of abduction, it leads to the Bayes theorem for confirming hypotheses, and naturally captures the theoretical virtue of quantitative parsimony. I address common criticisms against this approach, including how to assign probabilities to sentences, the problem of the catch-all hypothesis, and the problem of auxiliary hypotheses. Finally, I make a tentative argument that mathematical deduction fits in the same probabilistic framework as a deterministic limiting case.
# Contents

1 Introduction 1

2 Deduction 7
   2.1 Deductive Inference and Logic 7
   2.2 Mathematics = Science? 14
   2.3 Examples 17

3 Induction 29
   3.1 The Problem(s) of Induction 29
   3.2 Probability 32
      3.2.1 The Classical Theory 34
      3.2.2 The Logical Theory 36
      3.2.3 The Subjective Theory 39
      3.2.4 The Frequency Theory 42
      3.2.5 The Propensity Theory 44
      3.2.6 A Modified Propensity Theory 47
      3.2.7 Syntactic Elegance and Objective Probability 50
      3.2.8 A Paradox? 51
   3.3 Probabilistic Deduction and the Problem of the Catch-all Hypothesis 52
   3.4 Examples 54

4 Abduction 64
   4.1 Bayesian inference 65
      4.1.1 Quantitative Parsimony and the Problem of Auxiliary Hypotheses 67
4.2 The Error-statistics Approach ................................. 70
4.3 Examples ....................................................... 73

5 Conclusions ......................................................... 83
  5.1 Future Directions ............................................. 85

References .......................................................... 88
List of Tables

3.1 A summary of the different conceptions of probability throughout history, summarized from Gillies (2000). Each postulate the extra axioms listed in the second column. The third column states the metaphysical nature of probability, i.e., whether it is epistemic, ontological, or both. The fourth column states whether conditional probabilities are conceived by definition or as another axiom (or not at all). The fifth column shows the semantics of the probability theory, i.e., what is the object that receives a numerical value of probability.

3.2 Top table: probabilities of scientific hypotheses and conditional probabilities for observed astronomical phenomena given the hypotheses. Bottom table: probabilities of theorems calculated using Theorem 5 with the values above.

3.3 Top table: probabilities of hypotheses and conditional probabilities given the hypotheses related to Parkinson-like symptoms. Bottom table: probabilities of theorems calculated using Theorem 5 with the values above.

3.4 Top table: probabilities of hypotheses and conditional probabilities given the hypotheses related to the possibility of Bob cheating on his wife. Bottom table: probabilities of theorems calculated using Theorem 5 with the values above.
3.5 In this case we must split our hypotheses into postulated laws and the coefficients each law implies. Top table: prior probabilities on the laws and the conditional probabilities of the coefficients related to polynomials. Middle table: prior probabilities on the conjunctions of the laws and their coefficients, along with the likelihoods of observations of empirical data given each hypothesis. Bottom table: probabilities of theorems calculated using Theorem 5 with the values above.

4.1 Posterior and prior probabilities, along with likelihood and expectedness values for the case of Ptolemaic versus Copernican astronomy. Graphically represented in Figure 4·1.

4.2 Posterior and prior probabilities, along with likelihood and expectedness values for the case of diagnosing Parkinson-like symptoms. Graphically represented in Figure 4·2.

4.3 Posterior and prior probabilities, along with likelihood and expectedness values for an alternate case of diagnosing Parkinson-like symptoms. Graphically represented in Figure 4·3.

4.4 Posterior and prior probabilities, along with likelihood and expectedness values for the case of determining whether or not Bob is cheating on his wife. Graphically represented in Figure 4·4.

4.5 Posterior and prior probabilities, along with likelihood and expectedness values for the case of finding a polynomial curve to fit empirical data. Graphically represented in Figure 4·5.
List of Figures

1-1 Reality presents itself to us through observation. From these we construct mental models of some of its features. Two witnesses of the same observations may construct different models. I focus on the middle-top box, where we represent semantic concepts in a formal language, and then either deduce theorems or abduce a best hypothesis. Without the common language there can be no comparison of the models. 5

2-1 Deductive system related to the ZFC system of mathematics. There are ten axioms from which theorems are deduced, and from these theorems other theorems are deduced. The theorems (and not their negations) are deduced with absolute certainty given that the axioms (and not their negations) are known to be true with absolute certainty. 18

2-2 An arbitrary deductive system. Within a closed curve the possible sub-hypotheses or sub-theorems are mutually exclusive. 19

2-3 Deductive system with multiple hypotheses: Ptolemaic versus Copernican astronomy. 24

2-4 Deductive system with multiple hypotheses: Causes of Parkinson-like symptoms. “Chemicals in hard drugs” does not have a clear implication to whether the onset of symptoms occurs at a young or old age, hence no arrow to either theorem. 25
2-5 Deductive system with multiple hypotheses: Bob’s guilt or innocence with respect to cheating on his wife. There is no clear implication from the hypotheses to the theorems, hence no arrows. This case can only be examined in a probabilistic system. 26

2-6 Deductive system with multiple hypotheses: Fitting a polynomial of some order to observed data. Since this is a purely mathematical example, there are clear implications from each of the hypotheses to each of the theorems. The theorems involve observed empirical data that lie along the curves predicted by the different polynomials along with the different choices of coefficients. 27

2-7 Polynomials implied by the four different hypotheses of Figure 2-6. The curves are the predictions (theorems) that can be deduced from each hypothesis. 28

3-1 A graphical representation of the information in Table 3.2. 59
3-2 A graphical representation of the information in Table 3.3. 60
3-3 A graphical representation of the information in Table 3.4. 61
3-4 A graphical representation of the information in Table 3.5. 63

4-1 Posterior probabilities on the hypotheses for Ptolemaic versus Copernican astronomy (from Table 4.1). Red stars indicate the theorems that were observed. The deductive tree from Figure 3-1 is ‘pruned’ to remove all implications that do not lead to observations. 78

4-2 Posterior probabilities on the hypotheses for the diagnosis of Parkinson-like symptoms (from Table 4.2). Red stars indicate the theorems that were observed. The deductive tree from Figure 3-2 is ‘pruned’ to remove all implications that do not lead to observations. 79
4.3 Posterior probabilities on the hypotheses for the diagnosis of Parkinson-like symptoms (from Table 4.3). Red stars indicate the theorems that were observed. The deductive tree from Figure 3.2 is ‘pruned’ to remove all implications that do not lead to observations. 80

4.4 Posterior probabilities on the hypotheses that Bob is or isn’t cheating on his wife (from Table 4.4). Red stars indicate the theorems that were observed. The deductive tree from Figure 3.3 is ‘pruned’ to remove all implications that do not lead to observations. 81

4.5 Posterior probabilities on the hypotheses of which order polynomial and which coefficients best predict observed empirical data (from Table 4.5). Red stars indicate the theorems that were observed. The deductive tree from Figure 3.4 is ‘pruned’ to remove all implications that do not lead to observations. 82
List of Abbreviations

FOL ............ First Order Logic
FRPS ............ Falsifying Rule for Probability Statements
ZFC ............ Zermelo-Fraenkel Axioms with Axiom of Choice
Chapter 1

Introduction

Inference is how we obtain knowledge of the world around us; we do it almost constantly, sometimes without being consciously aware of it. The term can refer to deriving conclusions from premises we assume are true, such as when given that \( x = 5 \) and \( y = 2 \), we can infer that \( x + y = 7 \). It can also refer to the process of constructing a hypothesis that explains current observations and predicts new ones, such as when upon observing thousands of white swans, we infer that all swans are white. Inference can also be the process of choosing between several competing hypotheses that all explain observations, e.g., when we find the lawn wet in the morning, we can infer that it rained during the night. Although each of these examples is a case of inference in a general sense, there seem to be marked differences between them, and, in fact, philosophy has traditionally understood each as representing a different concept.

The first example is a mathematical deduction, which is a special case of deductive inference. Deduction is transformative; it does not create information, but rather brings out the (not necessarily apparent) implications of a set of premises through accepted rules of logic.

Our second example above is a case of inductive inference, and, unlike deductive inference, it is understood to be ampliative with respect to information; it seeks to take a finite number of observations and infer a general rule from them which can then be used to make an unlimited number of future predictions. The observation that all swans are white allows one to predict the color of the next swan to be encountered,
and the one after that, etc. Inductive inference essentially seeks to create a deductive system that logically implies the observations.

The third example above represents abductive inference, also known as inference-to-best-explanation. It seeks to find the single hypothesis from among many that is the most likely cause of observations. In the case of the wet lawn in the morning there are other possible explanations for the situation: a sprinkler system, morning dew, etc. The study of abductive inference seeks to understand how one chooses one of these over the others.

The relationship between these three categories, all of which legitimately can be called “inference”, has not been well understood. Historically, philosophers have tried to formulate inductive inference based on deductive principles, but this has been fraught with difficulties and may not be possible. Abductive inference seems to be very similar to inductive inference, since both attempt to find a hypothesis from which one can deduce observations; the difference being that abduction chooses from among several existing options, while induction generates hypotheses ex nihilo. That said, abduction also resembles deduction in the sense that it is possible to deduce a best hypothesis given existing observations by using the methods presented below in Chapter 4.

My first goal in this work is therefore to clarify these issues by constructing a unified framework for inference from which deductive, inductive, and abductive varieties can be seen as merely different aspects. I begin by presenting deductive inference, which requires a language (syntax), rule(s) of symbolic manipulation, and premises from which all theorems in the language can be deduced. Due to its well-understood structure, scope, and it’s use as the basis for mathematics, I limit myself to the language of First Order Logic (FOL) in this work. I next present inductive inference and,  

\[\text{Gillies (1996) gives a thorough account of attempts at representing inductive inference through a deductive framework in the context of artificial intelligence.}\]
along with it, the concept of probability. Probability is required because sentences that refer to real world examples of inference often have varying levels of uncertainty associated with them, and probability has historically been the most common way to quantify it. By assigning probabilities to premises, and conditional probabilities to theorems given that the premises are true, we are able to construct a probabilistic logic. Non-probabilistic deductive systems (e.g., mathematics) are then special cases of this general structure where probabilities of sentences are either 1 (certain) or 0 (impossible). For induction I also need to present the concept of observation — a theorem known to be completely true in the deductive system. Observations can “prune the deductive tree” of all implications and premises that are inconsistent with them. The theoretical virtue of simplicity also arises naturally in a probabilistic framework, and is directly related to setting prior probabilities on premises and likelihood values on implications. I finally present abduction, which in a probabilistic logic is equivalent to either calculating the Bayesian posterior probability, or else searching for a hypothesis that maximizes the likelihood.

My second goal in this work is to argue that the Bayesian viewpoint captures all the relevant aspects of abductive inference. Chapters 2 and 3 can then be seen as background material necessary for the discussion of abduction in Chapter 4. I examine several problems traditionally associated with Bayesian confirmation theory: First, I look at the question of which approach should be used to assign probabilities to sentences. In Section 3.2.6 I make a case for a Modified Propensity theory by drawing upon the work philosopher Donald Gillies. Second is the problem of the catch-all hypothesis, where we cannot assign a probability to the negation of a sentence when it amounts to an infinite number of other sentences. This prevents us from calculating probabilities of theorems, and as a consequence, calculating posterior probabilities. I do not offer a definitive solution to this problem, but I paint an approach that may
lead in the right direction, and point out an inconsistency in the previous work of other philosophers. Following this I look at the problem of auxiliary hypotheses, where extra hypotheses are inserted into an inference problem in order to avoid rejection of an irrationally preferred choice; I argue for a couple of criteria that can be used to solve this problem.

Four examples are used throughout this work to show the connections between deduction, induction, and abduction, as well as the scope of applicability of the Bayesian method. They are drawn from the philosophy of science, medicine, law, and statistics. For each of these examples I assign probabilities to hypotheses and implications, and from them I calculate probabilities of theorems in Chapter 3, and then posterior probabilities of hypotheses in Chapter 4. An important point to make is that I am not suggesting that all inference problems can be practically solved through formal symbolic manipulation in Bayesian inference; this has never been done in the philosophy of science and is probably not possible. Rather, I pose my framework as a way of viewing all inference problems in principle, which sometimes leads to practical quantitative algorithms, but otherwise only us gives a structure in which to discuss differing viewpoints. The numbers I assign to probabilities and likelihoods in Section 3.4 are therefore not intended to portray their true values; rather, they just illustrate that to the extent such probabilities can be assigned, posterior probabilities of hypotheses can be calculated in all the examples.

In stating what this work is about, it is also important to clarify what it is not about. In Figure 1-1 I try to capture what I consider to be the “big picture” of how we infer knowledge about reality. The figure pertains to two different scientific theories that explain observations. The bottom box represents all true sentences that could possibly be uttered about the nature of planetary motion. These truths are then manifested in every astronomical observation that humans make. These
Figure 1.1: Reality presents itself to us through observation. From these we construct mental models of some of its features. Two witnesses of the same observations may construct different models. I focus on the middle-top box, where we represent semantic concepts in a formal language, and then either deduce theorems or abduce a best hypothesis. Without the common language there can be no comparison of the models.
observations are represented by the images in the middle box which show data being collected. At the upper left and right we have two boxes representing the minds of two different people who have witnessed similar observations — in this case Ptolemy and Copernicus. Although both theories incorporate mathematical laws that accurately predict the observations, they are obviously different in the ontological reality they postulate. We thus have two scientific systems, one approximately true (Copernicus), the other completely false (Ptolemy), that can both make accurate predictions. The box separating the two minds represents language, which is used to symbolically represent their mental models, formally check for internal consistency, and compare with each other.

There are many deep philosophical problems represented in the simple diagram in Figure 1. For instance, how can we be sure that the mental model of any observer will ever approach the truth? Or even more skeptically, perhaps there is no bottom box, and no deep laws of nature to strive towards — how can we be sure? Another question involves whether there is any logical reason to believe that the mental objects in one mind can accurately reference the ontological reality, or be successfully translated to the other mind via language. Will “gravity” in one mind ever mean exactly the same thing in the other mind? These issues are interesting, but out of the scope of this work. I will take it as a given that there is indeed a reality that we are trying to approach with our mental models (hence I am a realist). I will also take it as a given that through communication humans are able to at least approximately transfer the mental objects from their minds to the minds of others — after all, these have been the assumptions of most of humanity throughout history and seem to work. I will instead focus solely on the top level of the diagram where formal symbolic manipulation can be used to compare different possible structures of reality to the observations in order to choose the most rational one.
Chapter 2

Deduction

In this chapter I will discuss the first of the three categories of inference, *deduction*. My discussion will be centered around formal logic. I will focus on its basic structure and limits, and show that both mathematics and science are represented in the same language, and from this viewpoint they can be seen as not being very different from each other.

I will conclude with examples of formal deductive systems taken from the philosophy of science, medicine, law, and statistics, all of which will be expounded upon in later chapters.

2.1 Deductive Inference and Logic

Deductive inference is translational; it begins with a list of premises in a language, and attempts to find useful conclusions that are consistent with, and derivable from them. If the premises are taken as self-evident and not reducible to other statements, they are called *axioms*. It is the method of mathematicians, who seek to tease out interesting theorems from the basic axioms of set theory.

Two essential properties of a correct deductive argument are

1. The premises must be true.

2. The conclusion must be an inescapable implication of the premises.

The second of these properties is known as *validity*. Deductions that obey these two
properties are called *sound* arguments. The accepted rule of deduction\(^1\), known as Modus Ponens, allows us to infer that a conclusion is true given that the argument is sound. In symbolic form it is

\[
\{H, H \rightarrow T\} \models T
\]  

(2.1)

where \(H\) represents a premise or hypothesis\(^2\), \(T\) represents an implied sentence (a theorem), \(\rightarrow\) is the usual symbol for if/then conditional\(^3\), and \(\models\) denotes that the consequent is a valid conclusion of the sentences in the antecedent. Based on such an argument, \(T\) can now be used as an additional premise that can be combined with other sentences to make further deductions.

A famous example of a sound argument has \(H\) being “Socrates is a man.” and \(H \rightarrow T\) being “If Socrates is a man, then he is mortal.”. The conclusion of this argument, via Modus Ponens, is therefore “Socrates is mortal.”. Arguments can be valid without being sound; substituting the words “Mickey Mouse is a man” for \(H\) in the deduction would be such an example. Arguments can also be invalid, yet with true premises, such as when we substitute the words “If Socrates is a man, then he can fly.” for \(H \rightarrow T\).

Representing sentences by symbols that may have a truth or falsity value, and combining them with logical connectors produces sentential logic, the simplest artificial logical language. It would seem that nothing more is needed to study deductive inference; one must only make sure that the premises are true and that the deductions follow logically. Experience has shown us, however, that this is not enough,

---

\(^1\)According to Enderton (2001), there are different ways one can develop a deductive calculus: one can either have a single rule of inference, Modus Ponens, and assume many logical axioms, or else assume many rules of inference and no logical axioms. Either approach amounts to the same thing.

\(^2\)I will use *premise* and *hypothesis* interchangeably. I will also use the terms *prediction* and *theorem* as synonyms.

\(^3\)I will assume the reader is familiar with sentential logic and its usual symbols: \(\rightarrow\) “if/then”, \(\leftrightarrow\) “if and only if”, \(\land\) “and”, \(\lor\) “or”, and \(\neg\) “not”. 
and that a finer granularity is sometimes needed to examine our arguments. Beyond sentential logic, other formal languages build upon its basic vocabulary by including operators, variables, functions, predicates, and constants. Most, but not all, systems hold to the law of excluded middle ($H \lor \neg H$) and the law of double negation elimination ($\neg\neg H \rightarrow H$). The most notable exception to this is intuitionistic logic (Moschovakis, 2014) formulated by Dutch mathematician L.E.G. Brouwer (van Atten, 2011), which was used to create the non-classical, constructionist axiomatic system of mathematics. Among all classical formal systems of logic, the one that has received the most attention in the last century is First Order Logic (FOL). It underlies the Zermelo-Fraenkel axioms of set theory with Axiom of Choice (ZFC), and thus is the basis of practically all modern mathematics (and therefore necessary for all physics and most other sciences and engineering disciplines).

Formal logical systems remove much of the redundancy of ad hoc languages\(^4\) and boil arguments down to their essential components, which we represent as symbols. These symbols help us keep track of all the premises and implications at the same time and organize our thoughts. Additionally, by connecting syntactic symbols to semantic concepts, and then applying Modus Ponens to the symbols, we may arrive at true statements that would never have been clear to us if we kept and manipulated all the information in our minds. A case in point for this would be ZFC. From its ten axioms\(^5\) almost all of modern mathematics can be derived as theorems. Although we can hold the meaning of each of its axioms in our minds with perfect clarity, there is no way to internally arrive at the truths of calculus, topology, or algebra; these theo-

\(^4\)It should be noted, however, that the artificial languages used in logic, like our ad-hoc languages, are also redundant. In sentential logic, for example, multiple basic connective symbols are used, when it can be shown that only one is necessary to allow a complete characterization of truth, e.g., either the ↑ (“nand”) or the ↓ (“nor”) symbol. This seems to show that the most concise language is not necessary the most useful. At the cost of a larger vocabulary and redundancy we can shorten the length of the sentences we write.

\(^5\)The usual formulation of ZFC uses ten axioms. It can be shown, however, that only eight of them are independent.
ries have required symbolic manipulation done on paper by countless mathematicians over many generations to deduce. Another good example of a useful deduction we could never perform in our minds is the computer simulation of a complex physical system. We can understand in our minds the premises (the laws of physics with boundary/initial conditions), but we cannot work out in our minds what the simulation will be. The computer essentially works out the deductive implications of the premises and gives us the simulation as a theorem of the deductive system. This capability of fast and error-free symbolic manipulation is why computers were created in the first place.

A final benefit of formal logical systems is that they can ascertain truth of theorems based solely on the syntactical form of the argument. Whereas some statements in a language are true only relative to the semantic structure being referred to (such as the axioms of ZFC are true relative to sets), there are also certain sentences that hold true regardless of what the symbols refer to; i.e., certain sentences are universally valid regardless of the semantic content. In sentential logic these are known as tautologies. In FOL they are called logical axioms, and include the tautologies (see Enderton (2001, page 112) for a list of all the logical axioms).

The question might be asked “Which formal logical language is correct?” Some are incompatible with one another (e.g., intuitionistic versus classical logic), and we must choose one or the other in representing a certain universe of discourse. Among the classical formal languages, the domain of applicability may be different, in which case the answer may depend on which semantic concepts are being discussed. For example, discussing the necessity, possibility, or provability of objects may require the operators used in modal logic (Garson, 2014), rather than those of FOL. Just as certain concepts are easier to state in one informal language rather than another, perhaps it is the same in formal languages. In his book discussing recent computational
efforts towards artificial intelligence Donald Gillies (Gillies, 1996, pp. 95-97) gives examples where certain deductive languages have been more or less fruitful, depending on the problem at hand. The logical languages examined by Gillies are attempts at automating inductive inference; they have a logical structure, but also include a “control” that allows a computer program to find law-like behavior in empirical data. Gillies finds that certain languages are better at expressing certain semantic concepts than others, and therefore better at performing inductive inference:

... it is ... appropriate to use a logic which draws reasonable conclusions ... 

What emerges is the idea that there is not a single universal logic, but that different logics may be appropriate in different contexts or problem-situations.

Clearly then classical logic is of very great importance, but it is not universal, because there are application areas for which standard mathematics is not the appropriate tool.

Thus logic is indeed empirical rather than a priori.

Examining whether Gillies is correct in this claim is beyond the scope of this work, and I refer the reader to Gillies’ book for further investigation. I will only say that I am currently uncertain as to whether logic is indeed a priori. Given the success of FOL at many different applications, and the dearth of applications for other formal logical languages, I still entertain the idea that it may indeed be the language of logic for expressing any and all deductive concepts. Due to this scope of applicability and its centrality in mathematics I will only use FOL for the remainder of this work and assume it is appropriate for expressing our examples.

Since FOL subsumes sentential logic, it keeps the sentence symbols of sentential logic as 0th-order predicates. I maintain that the following substitutions into 2.1 using
higher order predicates are needed to represent the hypotheses and implications used in science, mathematics, and statistics that I will discuss below:

\[
H \leftrightarrow (\exists c(c = c')) \land (\exists c \forall x \exists y Lcxy) \tag{2.2}
\]

and

\[
T \leftrightarrow (\exists x_1 \exists y_1(x_1 = x'_1 \land y_1 = y'_1) \land \cdots \land \exists x_N \exists y_N(x_N = x'_N \land y_N = y'_N)) \tag{2.3}
\]

This requires some explaining. Symbols in boldface signify an array of objects, e.g., \(c\) is the collection of parameters \(\{c_1, c_2, \ldots, c_M\}\). Symbols with a ′ after them are constants, i.e., fixed objects in our universe of discourse. The equality predicate \(c = c'\) in #1 says that there exists an array of parameters (the \(c\)) in our universe that equals an array of constant values (the \(c'\)). The three-place predicate \(Lcxy\) in #2 is a law\(^6\) that describes a functional input/output relationship between \(x\) (the independent variables) and \(y\) (the dependent variables). It has adjustable parameters \(c\), which might be initial conditions, boundary conditions, coefficients of a differential equation, coefficients of a polynomial, etc. The existential quantifier over \(c\) says that there exists a particular choice of \(c\) such that the law holds. The for-all quantifier over \(x\) says that this relationship holds for all the objects in the universe that \(x\) can refer to. The final existential quantifier then says that for each \(x\) there is a corresponding output \(y\). #3 then lists the input/output combinations that constitute our set of observations.

The relationship described here may arise in a number of settings, e.g., when fitting a polynomial curve through data, or perhaps when modeling planetary motion. In

\(^6\)The “law” may be a fundamental physical relationship, or just an empirical model that fits data. I am not addressing the problem of explanation versus modeling in this work, and therefore do not distinguishing between the two.
the former, \( c \) would be the coefficients of a polynomial of some order (higher degree polynomials would have more coefficients), \( x \) would be the independent axes on a graph, \( Lcxy \) would be the form of the polynomial, and \( y \) would be the value on the dependent axis. In the latter example, \( c \) would be the masses of bodies, \( x \) would be spatial coordinates, and \( Lcxy \) would be Newton’s laws relating masses at different locations to the forces \( y \) on other bodies. I will examine these particular sentences again in Section 3.2 in the context of probability in inductive inference.

Returning to the concept of soundness, it can be restated as follows\(^7\):

\[
\text{If } H \vdash T \text{ then } H \models T. \tag{2.4}
\]

\( \vdash \) represents a sequence of steps using \( H \) along with the logical axioms and Modus Ponens to establish theorems in the formal language. This statement essentially says that if you can start with correct premises and prove something, then it is a true implication.

There is a converse statement to 2.4 known as completeness, and it occupies an equally important place in the study of logic:

\[
\text{If } H \models T \text{ then } H \vdash T. \tag{2.5}
\]

Completeness says that if something is true, then you can prove it in your formal language. Whereas soundness can be proved quite simply with a few lines of written argument, a proof of completeness is not so simple, and first appeared for the language of FOL in Kurt Gödel’s doctoral dissertation in 1930.\(^8\) Later, in his famous incompleteness theorem, Gödel went on to show that when expressing certain semantic structures with their associated axioms in FOL we no longer have completeness.

\(^7\)Note that this is a statement in English, \textit{not} in the formal language. \( \vdash \) and \( \models \) are symbols in 	extit{English} used to express concepts of provability and validity, respectively, in the formal language.

\(^8\)For proofs of both 2.4 and 2.5 see Enderton (2001, pp.131-141).
This means that there will be sentences that are true in a formal system, but are unprovable from the premises, thus pointing out a fundamental limit of reasoning deductively about the structures. In can be shown, for example, that ZFC and all the theory of mathematics that it produces, is incomplete. This result shook the foundations of the mathematical community — putting an end to efforts by mathematical giants such as David Hilbert to prove that all theorems in a formal mathematical system (like ZFC) could eventually be proved, given the correct axioms.

Another desirable property of a deductive system is that it should be consistent, i.e., it should not be possible to prove both $T$ and its opposite $\neg T$ from the premises. Such an inconsistency implies any statement whatsoever in the system\(^9\). Simpler logical systems can sometimes prove whether or not their premises are consistent. ZFC and most others systems of interest, on the other hand, cannot. For these formal languages the question of consistency is semi-decidable: one can never be 100% sure that the deductive system is consistent, but one can be 100% sure that it is inconsistent upon encountering a contradiction. Although consistency is an absolute requirement in a formal language like mathematics, where we can only ascribe truth or falsity to statements, we can relax this constraint by assigning varying degrees of truth to statements with probability theory, which we will do below in Section 3.2.

2.2 Mathematics=Science?

In both mathematics and science, the premises are open to revision, and this is usually done when a internal contradiction is encountered. In mathematics it is all-or-nothing — a contradiction is absolute and implies that the premises must be changed. A historical example was the revision of Frege’s Axiom of Comprehension in light of Russell’s paradox (Irvine and Deutsch, 2014). The inconsistent result was that a set

\(^9\)This is because the law of contradiction implies any statement: $(T \land \neg T) \rightarrow Z$, for any Z.
both could, and couldn’t be a member of itself \((T \land \neg T)\). In science, the contradiction is relative — if it is strong enough, the premises should be changed. An example from physics is the failure of general relativity at small scales and the failure of quantum field theory at large scales. The theorems implied by each are contradictory, which has led the physics community to try to revise its axioms and create a unified theory.

It may also be the case that there is no contradiction, yet an observation cannot be deduced from any of the premises. If it can be proved that the sentence is independent of the current premises, it can be included as a new premise\(^{10}\). This was done in mathematics when the Axiom of Choice was incorporated with the original Zermelo-Fraenkel axioms.\(^{11}\)

If a theorem can’t be derived from the premises, it’s independence can’t be proved, and there is no obvious contradiction with the current system, then the sentence remains as a “conjecture” until one of those conditions is met. In mathematics this is the status of various mathematical sentences such as “Goldbach’s conjecture” and the “Riemann Hypothesis”, for example. In physics, Dark Energy has this status; it is not certain whether it will eventually encompass another law, or if it will ultimately be a consequence of existing principles. It should be noted, though, that sometimes the derivation of a theorem from accepted premises is only found after centuries of work — as in the case of Fermat’s Last Theorem. We therefore see that there is an empirical aspect to deductive systems. Whether dealing with mathematics or science, there seems to be a progress over time; as the number of observations increases, the chance increases of either finding a proof from the premises, or showing a contradiction (and thus revising the premises), or else proving independence (and thus including it

\(^{10}\)Gödel, and later Chaitin (2003) advocate this quasi-empirical approach for mathematics and view it as a way for mathematics to grow and increase its scope of applicability.

\(^{11}\)The negation of Axiom of Choice could have been added to the Zermelo-Fraenkel axioms, instead. It would have generated a different theory of mathematics, but one without many of the theorems that most mathematicians take for granted.
or it’s negation as a new premise).

These comparisons raise the question of whether science and mathematics are truly different, and if so, what are the differences. Differences related to consistency and contradictions mentioned above are only in a relative sense, where sentences in mathematics lie at the absolute ends of the spectrum of truth and falsity. A more plausible difference comes in the meaning of “observation” in each language. In the sciences this term refers to a sensory measurement of some phenomenon, either directly, or indirectly with other instruments. In FOL this concept was represented by $(\exists x \exists y (x = x' \land y = y'))$, which said that in nature we see a relationship holding between certain $x$’s and $y$’s that are measured. I will argue that the above-mentioned conjecture/hypothesis (i.e., a mathematical relationship that always seems to hold whenever tested mentally) is equivalent to a theorem being observed to be true. In the case of mathematics, however, these theorems along with input/output values are thought up and analyzed in our heads in order to observe whether the relationship holds. Thus it seems that the same logical structure holds for both mathematics and science. The only difference is the semantic structure being modeled and the interpretation of what it means to be an “observation”.

Assigning probabilities to sentences (Chapter 3, below) can hypothetically be done in mathematics as it is in science; just as our senses can be fooled when observing scientific phenomena, our mathematical intuition can be fooled when observing mathematical phenomena. Or, put another way, just as we can assign uncertainty to scientific premises and observations, so can we assign it to mathematical axioms and conjectures. I therefore argue that perhaps the only difference between the deductive systems of science and mathematics are the objects being discussed and the values of probabilities assigned the sentences. This, however, may lead to a paradox, which I discuss briefly in section 3.2.8, below.
2.3 Examples

Figure 2.1 gives a graphical representation of the structure of ZFC mathematics. There are ten hypotheses, which in this case are assumed to be irreducible axioms, from which all mathematics can be derived. The dashed arrows show theorems being derived from the axioms. Exactly which axioms and theorems are being used here is irrelevant; the point is to show what the deductive tree would look like in the case of set theoretic mathematics. The mutually exclusive choices for hypotheses in each closed curve are the axioms and their negations. If we assume each axiom (and not its negation) is true, then the implied theorems are also true. If we were to assume the negation of an axiom (e.g., the Axiom of Choice) we would have a different system of mathematics. The negation of an axiom would then not necessary imply either the previously implied theorem or its negation.

Figure 2.2 shows another deductive tree that illustrates a couple of differences. In this case each hypothesis is not constrained to have two possible choices (the hypothesis and its negation). Instead it may have multiple sub-hypotheses. What is important to note here is that within a closed curve, the choices of sub-hypotheses are mutually exclusive. $H_{3,1}$ or $H_{3,2}$ may be true, but not both; however, $H_{3,1}$ may be true simultaneous to $H_{1}$ being true (because they are independent of each other). For examples where there is only one hypothesis containing multiple sub-hypotheses, I will refer to the sub-hypotheses as ‘hypotheses’ for convenience and simplicity.

In Figure 2.3 through 2.6 we show a few specific examples of deductive systems. These examples will be revisited in both Chapters 3 and 4 where we will inductively assign probabilities to sentences and abduce best hypotheses, respectively. The collections of sub-hypotheses in these cases are exhaustive due to the inclusion of the

---

I will refer to mutually exclusive sentences within a closed curve as *sub-hypotheses* or *sub-theorems* when needed to differentiate it from the idea of a hypothesis being the entire set of sentences within a closed curve.
Figure 2.1: Deductive system related to the ZFC system of mathematics. There are ten axioms from which theorems are deduced, and from these theorems other theorems are deduced. The theorems (and not their negations) are deduced with absolute certainty given that the axioms (and not their negations) are known to be true with absolute certainty.
Figure 2.2: An arbitrary deductive system. Within a closed curve the possible sub-hypotheses or sub-theorems are mutually exclusive.
the “other” hypothesis (this is known as the catch-all hypothesis, which will be examined in detail in Section 3.3, below). Again, if a theorem is a clear implication of a hypothesis, a dashed arrow is drawn from the latter to the former. Arrows are lacking from the catch-all hypothesis to any of the theorems, since without defining what the catch-all actually means, it makes no sense for it to imply anything in particular.

The first example in Figure 2·3 comes from the history of astronomy, where we compare Ptolemaic versus Copernican theories. The Ptolemaic theory postulates that the Earth is the center of the universe with other planetary bodies revolving around it. Thus it implies no parallax in astronomical observations of stars or varying brightness of planets, but it does accurately predict planetary motion. Copernican theory, on the other hand, puts the Sun at the center of the universe; while it also accurately predicts planetary motion, it additionally implies parallax and varying brightness of planets throughout the year.

Figure 2·4 shows an example of a hypothetical deductive system drawn from neuroscience where a doctor is trying to diagnose the cause of Parkinson-like symptoms based on a couple of observed attributes of the patient. The possible hypotheses here are that the symptoms are either caused by a so-called Parkinson gene (the usual cause in young patients), a slow build-up of Lewy bodies in the Substantia Nigra area of the brain (the usual cause in elderly patients), or else the side effect of chemicals found in hard drugs used by addicts. In contrast to the last example, arrows do not go from all hypotheses to all theorems; e.g., an arrow is missing from the hypothesis “chemicals in hard drugs” to theorem $T_1$, because it does not imply onset of symptoms in either old or young patients. In a probabilistic setting we can assign probabilities and thus allow implication arrows, but in this purely deductive setting we cannot.

The third example in Figure 2·5 pertains to the guilt/innocence of someone accused of wrong-doing. Although the canonical example for such a situation is a legal
trial of an accused defendant, I have chosen a more informal example for illustration here. In my example a man named Bob is suspected of cheating on his wife and three observations, if observed to be true, would seem to give damning evidence for this conclusion\textsuperscript{13}. These observations are that someone observed him with another woman at a bar, that his wife found a receipt from a local jeweler that she knows nothing about, and that Bob came home from work with lipstick on his collar. These can each be thought of as a theorem implied by a hypothesis in the deductive system. Although one hypothesis is that Bob is guilty of cheating, in his defense he has offered up a complex counter-explanation for these observations in which there was some other guy at the bar who looked like him, that his mother visited him at work and got lipstick on his collar by kissing him, and that his watch broke, thus resulting the bill with the jeweler. In this situation either hypothesis can imply either a theorem or its negation; in fact, considering that it is unlikely Bob will be caught in this situation regardless of whether he is cheating or not, it is more likely that the negations would be observed. Thus in this example no arrows are drawn from hypotheses to theorems. This shows the limits of a purely deductive system; further analysis requires probabilities to be assigned to hypotheses and implications.

The fourth example in Figure 2.6 comes from the fields of statistics/inverse theory, and fits well in a deductive system due to its purely mathematical nature. It involves examining different choices of polynomials that fit empirical data points. For this example we make use of the notation from FOL presented earlier. Each hypothesis is the conjunction of a law (in this case a polynomial functional relationship) and particular values for the coefficients used in the law. On the left of Figure 2.6 we have the hypotheses governing the functional form. In this case we are limiting the possible hypotheses to either a first or second order polynomial. Given a choice of polynomial order, there is an implication to a set of coefficients. In this case we only

\textsuperscript{13}I have borrowed this particular example from Calcott (2011).
allow for two possible combinations of coefficients for either the first or second order polynomials. For $L_1$, the first pair of coefficients, $(1.15, -0.15)$ perfectly predicts the $x - y$ pairs $(1, 1)$ and $(3, 3.3)$, but not $(2, 1.3)$. The second pair of coefficients, $(0.3, 0.7)$ perfectly predicts $x - y$ pairs $(1, 1)$ and $(2, 1.3)$, but not $(3, 3.3)$. Turning to the second order polynomial, its triple of coefficients $(1, 1, -8.7)$ perfectly predicts only the $x - y$ pair $(3, 3.3)$, but not the others. The coefficients $(0.85, -2.25, 2.4)$ on the other hand perfectly predict all three of the $x - y$ pairs. The situation we have is that the first order polynomial for either choice of coefficients can only predict two out of three observations. The second order polynomial can predict one pair for one choice of coefficients, but all three observations for the other choice. These deductions can be seen in Figure 2·6 by where the dashed lines from each of the coefficient choices maps to in each theorem. For example, each choice of coefficients $c'$ for the linear polynomial goes through two out of three of the observations. For the second order polynomial, one choice of coefficients predicts all three of the data points in the theorems, while the other only predicts one of them.

Again, the hypotheses in this example are whether the polynomial is first or second order (along with the catch-all hypothesis which includes every other possible curve that can fit the data) in conjunction with particular value of the associated coefficients. Since we are dealing with mathematics, we have no wiggle room regarding which theorems are predicted by a given choice of hypothesis; each hypothesis implies an exact curve and either our observed data points will lie along it, or not. The totality of all theorems predicted by each hypothesis can be seen by the curves in the graph of Figure 2·7, along with where they each intersect the observed data points. This particular example was constructed such that the each choice of hypothesis will imply some or all of the observations.

As a purely deductive systems go, none of these examples is very illuminating.
With the exception of the polynomial fitting problem, they are very different from a system like ZFC, which had each theorem being deduced with 100% certainty, and thereby adding useful information to the existing body of mathematics. In these systems, on the other hand, hypotheses might not imply some theorems with certainty. However, by assigning appropriate probabilities to the hypotheses, and conditional probabilities to theorems given hypotheses, in Section 3.2, below, we will be able to examine how probable particular theorems are, and then in Section 4.1 we will also see how to abduce the optimal hypotheses.
Figure 2.3: Deductive system with multiple hypotheses: Ptolemaic versus Copernican astronomy.
Figure 2-4: Deductive system with multiple hypotheses: Causes of Parkinson-like symptoms. “Chemicals in hard drugs” does not have a clear implication to whether the onset of symptoms occurs at a young or old age, hence no arrow to either theorem.
Figure 2-5: Deductive system with multiple hypotheses: Bob’s guilt or innocence with respect to cheating on his wife. There is no clear implication from the hypotheses to the theorems, hence no arrows. This case can only be examined in a probabilistic system.
Figure 2.6: Deductive system with multiple hypotheses: Fitting a polynomial of some order to observed data. Since this is a purely mathematical example, there are clear implications from each of the hypotheses to each of the theorems. The theorems involve observed empirical data that lie along the curves predicted by the different polynomials along with the different choices of coefficients.
Figure 2.7: Polynomials implied by the four different hypotheses of Figure 2.6. The curves are the predictions (theorems) that can be deduced from each hypothesis.
Chapter 3

Induction

The second category of inference is induction. It deals with how we construct laws that explain current observations and predict new ones. The study of induction has been limited by the so-called problem of induction, which we will examine below. The idea of assigning probabilities to sentences springs naturally out of induction, and we examine different philosophical approaches to doing this.

Using Popper’s Propensity theory as a starting point, we combine probability with logic to make a probabilistic deductive system. This allows us to analyze the probability of deduced theorems, but also requires us to examine the problem of the ‘catch-all’ hypothesis.

Finally, we conclude by assigning probabilities to the hypotheses, along with conditional probabilities of theorems given the hypotheses, for the examples in the previous chapter, thus allowing us to calculate the unconditional probability of the theorems being deduced.

3.1 The Problem(s) of Induction

Inductive inference is the process of using limited observations to construct a general rule that can predict future observations. It underlies the scientific method, where a hypothesis is proposed to explain phenomena and then tested against future observations in order to be validated. Following Peter Lipton (Lipton, 2013), the problem of induction can be split into three related questions:
1. *How* does a person create a such explanatory law *ex nihilo*?

2. What justification do we have that the rule should predict future observations?

3. If there are many explanatory laws that fit observations, how do we choose the right one? (Underdetermination)

Addressing the first question, and using the notation of FOL from above, the traditional goal of induction is to start with

\[ \exists x_1 \exists y_1 (x_1 = x'_1 \land y_1 = y'_1) \land \cdots \land \exists x_N \exists y_N (x_N = x'_N \land y_N = y'_N) \]  

(3.1)

and arrive at

\[ (\exists c \forall x \exists y \text{ Lcxy}) \]  

(3.2)

via some mechanical procedure. This is at the heart of attempts to ground inductive inference in deductive inference. Francis Bacon, in particular, sought this kind of approach to induction — a set of “rules”, which if followed would lead to a law for predicting new observations (quoted from Gillies (1996, pg. 3)):

> But the course I propose for the discovery of sciences is such as leaves but little to the acuteness and strength of wits, but places all wits and understandings nearly on a level. For as in the drawing of a straight line or a perfect circle, much depends on the steadiness and practice of the hand, if it be done by aim of hand only, but if with the aid of rule or compass, little or nothing; so it is exactly with my plan.

This was a mistake, as the failure of many decades of effort to create a formal inductive system shows. Rather, to answer the first question above, I view this aspect of induction as the inspirational act of creating a deductive system such that it explains current observations, and hopefully predicts new ones to some degree of accuracy.
In this viewpoint, induction and deduction are intricately tied together. Whereas
deductive inference is the process of deriving implications from a set of premises
along with valid implications, inductive inference is the process of constructing the
deductive system in the first place. There is no rule book to follow on how this
construction is done — or, better said, if there is such a rule book, it does not come
out of deductive inference, but is instead a capability of the mind, brain, or soul that
is best left for psychologists, neuroscientists, or theologians to study. This was the
viewpoint of Karl Popper, who said (Popper, 2013, pg. 409):

The initial stage, the act of conceiving or inventing a theory, seems
to me neither to call for logical analysis nor be susceptible of it. The
question how it happens that a new idea occurs to a man—whether it is a
musical theme, a dramatic conflict, or a scientific theory—may be of great
interest to empirical psychology; but it is irrelevant to the logical analysis
of scientific knowledge.

Whichever is the source of inductive inspiration, it is beyond the scope of my work
here. It is sufficient to say only that it can be done, and our collective experience
has shown humans able to come up with ingenious explanations that stand up to
experimental verification.

The second question above is closely related to the first: Once we have a law that
explains our current observations, what justification do we have that it will predict
future ones? A classic example of this problem is that upon observing 1000 white
swans we infer that the next swan will be white. Obviously no amount of white swans
will ever guarantee the conclusion. David Hume observed that if someone attempts to
justify induction by the fact that it has tended to work in the past, they are justifying
induction by induction itself, thus leading to a circular argument. Hume’s skeptical
conclusion was that there is no justification for it.
My only insight is that in asking the question about the future in view of the current uncertainty in our proposed law, it seems that the word *probability* naturally arises. With probability we don’t have to justify that the law will predict with complete certainty, but only that it will work most of the time. Although this seems to give us a little wiggle room, it really only pushes Hume’s skeptical argument one level back — instead of justifying a proposed inductive rule, we instead must justify the probabilities we assign to sentences in a deductive logic. We will examine current methods for doing this in the following section.

The final question above is closely related to the second, and stems from *underdetermination*, where we have many theories that can explain current observations equally well. In order to distinguish between the different hypotheses we must have a criterion for saying that one is better than the others. In my view this question is not one of induction, but rather lies in the domain of abductive inference, where we seek the best explanation for a set of observations; it will be examined in Chapter 4.

### 3.2 Probability

As mentioned in Chapter 2, a problem with formal deductive systems is that they cannot handle inexactness and uncertainty. Either a sentence is implied by the premises and implications, or it is not. This all-or-nothing viewpoint is limiting, but does not have to be synonymous with deductive inference. The concept of probability allows us to assign degrees of belief to premises, and conditional degrees of belief to theorems given a premise, in our deductive language, thus generalizing the deterministic concepts of certainty (probability 1) and impossibility (probability 0) to any fraction between 0 and 1.

In his book *Philosophical Theories of Probability* (Gillies, 2000), Donald Gillies gives an excellent overview of the topic from both a historical and theoretical per-
spective. According to him, there have been five proposed theories of probability: Classical, Logical, Subjective, Frequentist, and Propensity; to these he adds his own Intersubjective and Pluralistic theories. The key distinctions between them all are:

1. The mathematical axioms that define the theory.

2. The semantic object being referenced by the axioms (i.e., events, attributes, sentences, sets, etc.).

3. The justification for the values of probability being assigned to semantic objects.

4. Whether probability is an epistemic notion or an ontological one; i.e., whether the assigned values are due to human ignorance, or are a fundamental stochastic property of nature.

5. The status of conditional probability.

Regarding the first item, the usual approach in the applied sciences for developing a theory of probability follows that of the famous Russian mathematician Andrey Kolmogorov, who postulated three axioms (Hjek, 2012):

**Axiom 1 (Kolmogorov)** \[0 \leq P(A) \text{ for all } A \in F.\]

**Axiom 2 (Kolmogorov)** \[P(\Omega) = 1.\]

**Axiom 3 (Kolmogorov)** \[P(A \cup B) = P(A) + P(B).\]

For Kolmogorov the semantic objects are sets, with \(A\) and \(B\) being arbitrary sets, \(\Omega\) being the ‘universal set’, \(F\) are subsets of \(\Omega\), and \(\cup\) signifies the union of sets. Kolmogorov strengthened Axiom 3 to include countably infinite sets with the Axiom of Countable Additivity:

**Axiom 3’ (Kolmogorov)** If \(A_1, A_2, A_3, \ldots\) is a countably infinite sequence of disjoint sets, each of which is an element of \(F\), then \[P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n).\]

Although the semantics differ in different probability theories, these axioms are a minimal group shared by all of the them, and are thus used as a reference point by
Table 3.1: A summary of the different conceptions of probability throughout history, summarized from Gillies (2000). Each postulate the extra axioms listed in the second column. The third column states the metaphysical nature of probability, i.e., whether it is epistemic, ontological, or both. The fourth column states whether conditional probabilities are conceived by definition or as another axiom (or not at all). The fifth column shows the semantics of the probability theory, i.e., what is the object that receives a numerical value of probability.

Gillies. It is interesting that he does not explicitly put Kolmogorov’s view into any of his philosophical categories. My assumption is that he instead views it as a utilitarian approach that does not go deeply into its own philosophical foundations, but rather has shown itself useful to the applied community. Following Gillies, I will cover the essential properties of each theory in light of Kolmogorov’s Axioms, focusing on the five core philosophical distinctions listed above, four of which I summarize in Table 3.1.¹ For more detail I recommend the reader to Gillies’ book.

3.2.1 The Classical Theory

Up until the 1900’s, there was a common belief in universal determinism. This concept is most clearly illustrated by the concept of Laplace’s Demon who, given the initial conditions of reality in all its detail, could predict the state of the universe until

¹I do not put the Intersubjective or Pluralist theories in the table; the former because it shares the same properties as the Subjective theory, the latter because it is an amalgamation of the other theories.
the end of time. There was certainty that the laws of science had swept away all superstition and now allowed us to predict the motion of objects, from particles to planets, or any other phenomena we desired, with perfect accuracy, if only we have the correct initial/boundary conditions. With this background it is obvious that the concept of probability could only be an epistemic one, representing uncertainty in the mind of an observer of an event that theoretically could be known with certainty. Quoting Pierre-Simon Laplace from (Gillies, 2000, pg. 17):

> The curve described by a simple molecule of air or vapor is regulated in a manner just as certain as the planetary orbits; the only difference between them is that which comes from our ignorance.

Probability is relative, in part to this ignorance, in part to our knowledge. We know that of three or greater number of events a single one ought to occur; but nothing induces us to believe that one of them will occur rather than the others. In this state of indecision, it is impossible for us to announce their occurrence with certainty.

Implicit in this statement is the belief that given our state of ignorance, we should assign equal probability to each of the possible outcomes of a probabilistic scenario — the so-called uniform distribution. Probability is then the ratio of the number of specified events to the total number of possible events. A simple example of this, given by Gillies, is that of calculating the probability of an odd numbered side appearing in the roll of a six-sided die. The number of specified events is three, the total number of possible events is six (one for each side that can come up). Thus the probability is \( \frac{3}{6} = \frac{1}{2} \). This view of probability persisted into the early 1900’s; even the famous Russian mathematician, Andrei Markov, held to the classical view in his book on the subject, published in 1912. This view, however, could not address situations such the following, quoted by Richard Von Mises (Gillies, 2000, page 18):
But how are we to deal with the problem of a biased die by means of a theory which knows only probability based on a number of equally likely results?

It should be noted that Laplace and others were aware of this problem; indeed, Laplace examined the case of the biased coin, which seems to imply the existence of non-uniform objective probabilities, but for whatever reason this did not lead him to change his philosophical viewpoint.

The Classical theory also made no mention nor provided a definition of conditional probabilities.

3.2.2 The Logical Theory

The logical view of probability came predominantly from Cambridge and to a lesser extent from the Vienna Circle in the years before World War I. The name best associated with it is that of John Maynard Keynes. He identified probability with a degree of partial entailment (implication). In this view all probabilities are conditional, and only have meaning in relation to background information. Quoting Keynes from (Gillies, 2000, page 30):

No proposition is in itself either probable or improbable, just as no place can be intrinsically distant; and the probability of the same statement varies with the evidence presented, which is, as it were, its origin of reference.

The semantic objects that receive a probability in this theory are propositions, i.e., sentences.

Keynes further associates the ‘partial degree of entailment’ with the ‘degree of rational belief’, where the rational belief is predicated on the subject’s collective experience. He held that probabilities had a Platonic reality, and hence a true value.
Rational people would all assign the same true value of probability to a statement, but he also held that not all people are rational. Thus probability in his view was both ontological (in that there was a true value), and epistemic, since not all observers might have that true value in mind.

Regarding the assignment of probability by rational intuition, Gillies (2000, page 53) says

Frege, one of the greatest logicians of all time, was led by his logical intuition to support the so-called axiom of comprehension, from which Russell’s paradox follows in a few lines. Moreover, he had companions in this error as distinguished as Dedekind and Peano. Hilbert and Brouwer were two of the greatest mathematicians of the twentieth century. Yet Hilbert’s logical intuition informed him that the Law of Excluded Middle was valid in mathematics, and Brouwer’s that it was not valid. All this indicates that logical intuition is not to be greatly trusted in the deductive case, and so hardly at all as regards to inductive inferences.

Karl Popper also argues against Keynes’ method. Quoting from Gillies (2000, page 31):

... Popper’s argument continues, although the degree to which finite evidence partially entails a universal generalisation is zero, it may nonetheless be possible to have a non-zero degree of rational belief in a universal generalisation given finite evidence. Indeed this is often the case when we entertain some finite degree of rational belief in a scientific theory. So, Popper concludes, we should not identify degree of partial entailment with degree of rational belief. Popper accepts a logical interpretation of probability where probability is identified with degree of partial entailment, but, since these degrees of partial entailment are no longer degrees
of rational belief, his theory differs from that of Keynes. Popper identifies
degree of rational belief with what he calls ‘degree of corroboration’....

Related to this was Popper’s belief that there should be some objective way of altering
probabilities based on observations. Such an idea was completely lacking in Keynes
theory. Popper (2010, page 398) says:

... *we may learn from experience more and more about universal laws
without ever increasing their probability*; that we may test and corrobo-
rate some of them better and better, thereby increasing their degree of
corroboration without altering their probability whose value remains zero.

What I believe Popper is alluding to is that, regardless of what probabilities may be
assigned to statements, the results of empirical observations should be able to refute,
or falsify, the rationally chosen probability value. I will return to this issue in the
Frequency and Propensity theories, below.

As part of his theory, Keynes postulated a ‘Principle of Indifference’ as a way to
assign probability values to cases where we have no known reason to give one state-
ment more probability than another. This principle would equally divide probabilities
between all the possible outcomes, again leading to the uniform distribution, similar
to the Classical Theory. The difference being that in Keynes’ view this rule could
be superseded by the rational person if they believed that the sentences in question
should have a non-uniform distribution of probabilities (say, if they knew a die was
loaded). This principle, however, led to paradoxes when it was noticed that if a uni-
form distribution of probability was assigned to a continuous variable and then the
problem were re-parameterized, in the new parameterization the distribution would
no longer be uniform. The obvious question is then ‘Which parameterization is cor-
rect?’, but there is no straightforward answer to that question. Thus in cases of
continuous variables there is no objectively correct way to assign probability; a bias
is introduced by rational agents who may think they are being as unbiased as possible by choosing the uniform distribution. For examples of paradoxes arising from the Principle of Indifference, see, e.g., (Gillies, 2000, pp. 37-42).

Keynes also held that not all probabilities have numerical values. Sometimes we can only give comparative rankings of sentences. For instance, we can’t assign a probability to the sentence “The USA will go to war with Iran in the next 10 years.”, or the statement “The USA will go to war with Canada in the next 10 years.”, yet most people would rank the probability of the former as being higher than the latter.

### 3.2.3 The Subjective Theory

The unsuccessful resolution of the paradoxes arising from the Principle of Indifference led investigators to seek an alternative formulation of probability, which they found in the Subjectivist viewpoint. The original proponents of this view were Frank Ramsey in Cambridge and Bruno de Finetti in Italy, who discovered it independently. In this view we abandon the idea that all rational individuals will have access to some Platonic reality regarding the values of probability. Rather, each person has their own subjective value which they are quite entitled to assign to an event — the semantic object referred to by the theory. The main argument for how an individual should assign probabilities in the Subjective viewpoint revolves around the concept of betting quotients (Gillies, 2000, pp. 55-58), i.e., a person should assign their probabilities as if they were going to bet according to the odds they implied. Quoting Ramsey in (Gillies, 2000, pg. 54):

> The old-established way of measuring a person’s belief is to propose a bet, and see what are the lowest odds which he will accept. This method I regard as fundamentally sound.
The subjectivists were able to show that their approach satisfies Kolmogorov’s axioms. They did this through the idea of ‘coherence’, which essentially says that when assigning probabilities, a rational agent won’t do it in such a way that they are guaranteed to lose, i.e., a Dutch book cannot be made against them. The Ramsey-De Finetti Theorem then says that a set of betting quotients is coherent (i.e., doesn’t allow a Dutch book) if and only if it satisfies Kolmogorov’s axioms. Gillies (2000, pp. 59-63) provides the proof of this theorem.

A crucial difference between the Subjectivist view and Kolmogorov’s is how to interpret conditional probabilities. Kolmogorov viewed them as being defined as the ratio between probabilities:

\[ P(B|A) := \frac{P(A \cap B)}{P(A)} \quad (3.3) \]

Where \( \cap \) signifies the intersection of two sets.

Ramsey and de Finetti, on the other hand, have conditional probability as being a fundamental notion requiring it’s own axiom:

\[ P(A \cap B) = P(A)P(B|A) \quad (3.4) \]

A further technical difference is whether or not probabilities can, and should, be applied to countably infinite events. Kolmogorov said ‘yes’ (which required his additional axiom for handling such a case), while de Finetti, citing the fact that all bets are made with finite number of options, said ‘no’. This difference, however, is more a difference between de Finetti and Kolmogorov, rather than one between Kolmogorov and subjectivists in general, as there are many subjectivists who do, in fact, allow for applying probabilities to countable sets.

Whereas the Logical approach to probability associated it with degree of rational belief and its access to a Platonic reality, the Subjective theory associated it with
degree of belief — period. The Subjective view then was epistemic, which set it apart from the Logical view that allowed probability to be objective. Although the Subjectivist viewpoint seems to have solid grounding via the Ramsey-De Finetti Theorem, it does not account for cases where there actually seem to be objective probabilities such as, e.g., the probability of fair coin coming up as ‘heads’ half the time. De Finetti’s answer to this would be that we may assign any degree of subjective probability we wish; with each flip of the coin the subjective prior will then be modified into a new probability via Bayes rule (Chapter 4, below), and rather quickly the effect of whichever subjective probability was originally used would disappear.

There is a subtle problem here though, that Gillies goes to great lengths to point out: De Finetti’s Subjective theory relies on an idea of exchangeability, the subjectivist’s notion that any events being considered are independent of each other. Upon observation, the subjectivist can only modify prior probabilities into posterior ones by Bayes rule, but can never modify the hypothesis that the sequence of events is independent. Thus, if probabilities of the events are indeed dependent, the subjectivist will never arrive at accurate values of probability, regardless of how much Bayesian conditionalization goes on. Gillies (2000, pg. 74 and 80) says

In De Finetti’s scheme, we do not try to test or refute our prior probabilities . . . , we simply change them into posterior probabilities . . . by Bayesian conditionalisation.

Yet if we assume exchangeability a priori when the sequence of events is in reality dependent, no amount of modifying our prior probabilities . . . to posterior probabilities . . . by Bayesian conditionalisation will produce probabilities which accord with the real situation . . . Unless we know that the events are objectively independent, we have no guarantee that the use of exchangeability will lead to reasonable results.
De Finetti’s answer to the problem worked only because he was talking about events that truly are independent, such as the coin flipping experiment.

What we are seeing here is that the Subjective theory has a similar problem to that of the Logical theory with respect to the interaction between assigned probabilities and empirical observations. Whereas there was no mechanism for modifying probabilities in the Logical theory, the Subjective method does have Bayes rule. However, that is not enough. There is a deeper hypothesis hard-wired into the Subjective theory that cannot be changed, even if it would allow for better conditionalization. This situation would be analogous to the inductive problem of creating a hypothesis to explain observations where we are not free to choose the best one. This issue will appear again in the Frequency theory, but does not have a resolution until we examine the Propensity theory, below.

In his book, Gillies also covers an Intersubjective view of probability, which is almost identical to the Subjective view — the only difference being that the ‘subject’ who defines probabilities via betting quotients in the Intersubjective theory is a group, not an individual. All other properties and axioms remain the same. If the Subjective viewpoint is plausible, then so is the Intersubjective one. One could view Intersubjective probabilities as being part of a paradigm, in the Kuhnian sense. Since there are no further differences between the Subjective and Intersubjective view, I will not cover it further.

3.2.4 The Frequency Theory

The Frequentist approach is most closely associated in philosophy with Hans Reichenbach and Richard Von Mises of the Vienna Circle. It’s distinguishing mark is the fact that probabilities are obtained from collections of events that have attributes in common. For instance, in flipping a coin, the attributes are that 1) A coin is flipped, and 2) It came up ‘heads’. One can then compare ratios of events where
both 1 and 2 occurred to those where only 1 occurred. In this way the Frequentist approach carries within it what Gillies calls the ‘Axiom of Convergence’, which says that as the number of samples goes to infinity, the limit of the ratio of attributes exists. Von Mises’ justification for this axiom comes from the empirical sciences, where exactly this thing is done, i.e., scientific theories are (hopefully) true generalizations to infinite collections that can be obtained from finite samples. We see here another hint that the problem of assigning probability is close to, if not equivalent with, the problem of induction.

Frequentist probabilities are obtained from a finite number of samples for a finite number of attributes. Therefore the theory does not need Kolmogorov’s axiom of Countable Additivity, since it would imply human experience with infinite numbers of attributes. The Frequency theory therefore deals with infinite collections samples only in a hypothetical way, i.e., that if given an infinite number of samples of attributes, there would be convergence to a probability value.

In order to make a coherent theory Von Mises’ required an additional axiom called the Axiom of Randomness, which has no counterpart in Kolmogorov’s system. It states that there can be no order in a collection of samples from which we calculate a ratio; i.e., by selecting a subset of the samples the ratio should not change very much. Otherwise, if samples are not random then there exists some dependence of one sample on another, and selecting dependent samples can bias the resulting empirical ratio. Randomness of the sequence of samples therefore implies their independence and our ability to calculate frequencies. The Axiom of Randomness then plays the same role as the principle of exchangeability did in the Subjective theory.

Probability in the Frequentist view is always conditional, similar to the Logical viewpoint. It is conditional precisely on the existence of the collection of samples from which it obtains its value; i.e., values of conditional probability will be obtained
by an empirical ratio approximating Equation 3.3. Thus we must limit ourselves to the case where \( P(A) = \lim_{n \to \infty} \frac{n(A)}{n} \neq 0 \) in order to make such a calculation. This, however, leads to problems in situations where we may wish to ascribe probabilities when not even a single sample exists, e.g., the probability of the USA going to war with Canada in the next 10 years. Regarding this point, Von Mises says (Gillies, 2000):

‘The probability of winning a battle’, for instance, has no place in our theory of probability, because we cannot think of a collective to which it belongs. The theory of probability cannot be applied to this problem any more than the physical concept of work can be applied to the calculation of the ‘work’ done by an actor in reciting his part in a play.

3.2.5 The Propensity Theory

The aim of the Propensity theory was to overcome the main philosophical problem associated with the Frequency theory, namely that of assigning a probability to a singular event. Karl Popper, the original proponent, saw that the change needed to correct the Frequency theory was to abandon the idea of a collective from which to draw inference, and instead speak about a set of generating conditions that had a propensity or disposition to produce the collective.

I stated above that the problem of justification for induction could be kicked back to the problem of justifying the assignment of probabilities. The Propensity theory’s answer to this problem is paradoxically the same answer as to how someone creates a hypothesis ex nihilo to explain data — which is to say, there is no answer — or at least not one that comes from any rational principle I am aware of. In the Propensity theory the probability assignment must be created ex nihilo, also. Again, I offer no insight into how this is done, but once it is done we can then test how well the probabilities predict. The way to interpret this is that as we obtain more empirical
samples we can calculate ratios as we do with Frequency theory. To the extent that
the predictions corroborate the assigned probabilities, we retain them. To the extent
they don’t, they are falsified, and we must come up with another assignment. Gillies
agrees with this assessment, saying (Gillies, 2000, page 160):

The propensity theory . . . does not offer an explicit definition of prob-
ability from which the [Kolmogorov] axioms can be derived. It rather
regards probability as implicitly characterised by a set of axioms which
are designed to provide a mathematical theory of observed random phe-
nomena. The axioms are justified by showing that from them results can
be derived which are in agreement with observation.

Gillies advocates an absolute test for falsification, which he calls the ‘Falsifying Rule
for Probability Statements’ (FRPS). It basically amounts to something like a $\chi^2$-type
test, where if an observation occurs whose probability under the assumed probability
assignment is very small, the assignment is falsified. The question is then ‘How small
is too small?’, which has no absolute answer. I return to this idea in the context of
Deborah Mayo’s “error-statistics” approach below in Section 4.2.

From the perspective of someone familiar with statistics, what the Propensity
method is saying is that one must come up with some kind of parametric model of
possible outcomes of an experiment that generalizes beyond the current collection of
empirical samples. Then one continues to collect empirical samples and as long as
empirical ratios approximately match the parametric model, the model is retained.
If they don’t, another model must replace it. In this way we see that the method of
assignment of probabilities inherent in the Propensity theory is exactly the scientific
method: propose a hypothesis (in this case the hypothesis is the assignment of prob-
abilities), collect data, and see if the data falsifies the hypothesis (where falsification
is in a statistical sense of comparing empirical ratios with the probabilities).
Returning to the Frequency theory, we see there that it made assignments based of past experience, i.e., by taking empirical samples from the past and calculating ratios related to a particular attribute. This seems to me to be good for explanation, i.e., finding the probability that is most consistent with the empirical ratios that we used to derive it, but it is not predictive. The Propensity theory then makes the inductive leap to generalize and predict ratios for unobserved attributes.

The Propensity theory also subsumes the Subjective theory in the sense that the inspirational creative act of assigning probabilities is subjective — different people will do it in different ways. The difference is that the Propensity theory says that there exists a true assignment that best explains all possible future observations, whereas the Subjective theory says such a thing does not exist.

The Propensity theory also handles the problems of independence/exchangeability in both the Subjective and Frequency theories by re-defining what constitutes a sample as part of its assignment of probabilities. Practically speaking, this means that if we have a sequence of observations where each group of five is assumed to be probabilistically dependent, we can account for this by assuming a probabilistic model that has exactly that characteristic. Then, when making observations we must break up the sequence accordingly and treat each group of five as a separate ‘sample’ for purposes of calculating empirical statistics. If the assumption was wrong — say, there are actually groups of six dependent samples at a time — we can make alternate proposals and compare to see which one best fits the empirical data. This solution, we recall, could not be done with either the Subjective or Frequency theories. Gillies calls this requirement of only using independent collections the ‘Axiom of Independent Repetitions’. It seems straightforward — almost trivial — so much so that I would personally not include it as an axiom, but rather as simple common sense.

Finally, the Axiom of Countable Additivity is also justified by the Propensity
theory, since we can propose any number of probabilistic objects we wish, as long as the resulting theory is supported by observations.

Summarizing its superior properties, Gillies makes it clear that he sees the Propensity theory as the best among the different views of probability (Gillies, 2000, pg. 184):

The propensity theory ... explains conceptual innovation in the natural sciences better than Von Mises' operationalism; the propensity theory eliminates all the problems about infinite collectives, and, by introducing a falsifying rule for probability statements, gives an account of the relations between probability and frequency which agrees very well with the standard statistical practice; the propensity theory eliminates Von Mises' introduction of the two separate concepts of randomness and independence by reducing both to independence; the propensity theory fits in better with the Kolmogorov axioms and the modern mathematical approach to probability using measure theory, since it allows probability to be introduced as an undefined concept; and so on. Taking all these points together, I think we can definitely say that the propensity theory has superseded the frequency theory.

I find it hard to disagree with Gillies on his points, and I see the Propensity theory as the best basis for assigning probabilities.

3.2.6 A Modified Propensity Theory

Gillies maintains that only the Subjective, Propensity, and Intersubjective theories are viable (which I agree with). He claims that since two are epistemic theories and the other ontological, they need not contradict with each other and could be applied in different situations. He proposes this as the basis for a Pluralistic theory in which he
uses different theories in different contexts. Since the Propensity theory can account for every aspect of probability necessary for practical application, I see the other two theories as being superfluous. I do not understand why Gillies sees the need to retain them and mesh them into a more complicated Pluralist theory. I therefore depart from Gillies, and I view the Propensity theory as being appropriate to handle any probabilistic situation (whereas the other theories are lacking in one respect or another). The Propensity theory has the structure of a science, where its hypotheses (the probability values) are empirically tested and can be falsified. Subjectivity enters into it through the subjective choice that individuals can propose for the probabilities — either individually, or as a group. Frequencies enter through empirical testing. How to actually assign probabilities will depend on the application; when there appear to be objective probabilities, they should be used. If the probabilities are related to an empirical phenomenon, examining data for statistical patterns might be a good place to start. If they involve a naturally occurring range of possibilities (two sides of a coin, six sides of dice, etc.), a uniform distribution on the possibilities may be a good place to start. If they involve an occurrence with no empirical data or obvious choices, intuitive insight from similar past experience may point the way. The hypotheses and probabilities generated in this way will stand until they can be shown to be inferior to another competing set of hypotheses and/or probabilities.

Although this method seems to provide a rational approach that covers all aspects of probability, it assigns probability to ‘events’, which I believe is incorrect. Ascribing it to sentences, as the Keynes’ Logical theory does, seems more general than ascribing it to either events, attributes, or sets, and thereby makes probability a context-free theory. There also seem to be cases where a sentence can have a legitimate probability assigned to it, even though it doesn’t classify as an ‘event’ or a ‘set’, e.g., ‘the probability of the existence of God’, ‘the probability of the law of excluded middle’, ‘the probability of a tautology’, ‘the probability of a contradiction’, or ‘the probability of the axioms of ZFC’. These are not measurable occurrences in space-time (i.e., an ‘event’) or collections of things (i.e., a ‘set’), but we could place a probability
on the statements, if we wished. Thus in my view it is more appropriate to assign a probability to a collection of symbols that refers to an object in our mind, which then (hopefully) refers to an object in reality. Assigning probabilities to sentences allows one to apply it to a logical deductive system, thus resulting in a probabilistic logic. I develop this here, beginning with a re-statement of Kolmogorov’s axioms, and having conditional probability included as an additional axiom:

**Axiom 1**  Every probability is a real number between 0 and 1: $0 \leq P(A) \leq 1$.

**Axiom 2**  If $A$ is a necessary truth (i.e., a tautology), then $P(A) = 1$.

**Axiom 3**  If $A$ and $B$ are mutually exclusive (that is, $\neg(A \land B)$ is a tautology), then $P(A \lor B) = P(A) + P(B)$. *(Special Addition Rule)*

**Axiom 4**  **Conditional Probability**

$$P(H \land T) = P(H)P(T|H)$$

$A$ and $B$ here are sentences.

From these axioms we can deduce the following theorems:

**Theorem 1:**  **Negation Rule**

$$P(\neg A) = 1 - P(A)$$

**Theorem 2:**  **Equivalence Rule**

If $A$ and $B$ are logically equivalent, then $P(A) = P(B)$.

---

I originally investigated an alternate way of arriving at conditional probabilities. I made the assumption that a premise is probabilistically independent of an implication involving the premise, which would allow for the following relationship:

$$P(H \land (H \rightarrow T)) = P(H)P(H \rightarrow T)$$

Equation (3.5)

$$= P(H \land T)$$

Equation (3.6)

William Rodi, one of the readers of this thesis, pointed out that this would lead to a paradox if $\neg H$ were substituted for $T$; this would say that $P(H)P(\neg H) = P(H \land \neg H) = 0$, thus implying that either $P(H) = 0$ or $P(\neg H) = 0$. This, in turn, would render probability theory intractable. The conclusion seems to be that the probability of a hypothesis is not independent of the probability of an implication involving it, and therefore $P(H \rightarrow T) \neq P(T|H)$, i.e., the probability of a conditional is not equal to the conditional probability.

$A \lor \neg A$ is a tautology, thus we have that $1 = P(A \lor \neg A) = P(A) + P(\neg A)$. Rearranging gives the theorem.

$A \leftrightarrow B$ implies that $A \lor \neg B$ is a tautology, so $P(A \lor \neg B) = 1$. Since $A$ and $\neg B$ are contradictory statements we have that $P(A \lor \neg B) = P(A) + P(\neg B)$. From this we have that $P(A) + 1 - P(B) = 1$, and thus that $P(A) = P(B)$. 

Theorem 3: \textit{General Addition Rule}^{5}

\[ P(A \lor B) = P(A) + P(B) - P(A \land B) \]

3.2.7 \textbf{Syntactic Elegance and Objective Probability}

Although I believe my Modified Propensity theory is a valid approach to probability, there is one other viewpoint that may offer an improvement, and should be mentioned. It involves the idea that probabilities can be objectively assigned to sentences based on their syntax. The shorter the syntax, the more ‘elegant’ the description, and hence the more probable the sentence is. Elegance is then one of two conceptions of the theoretical virtue of simplicity found in the literature (the other, ‘parsimony’, is examined in Section 4.1.1, below). The use of elegance has appeared independently in a several different areas of mathematics and computer science in the past century — most notably in the Minimum Description Length Principle (Grünwald, 2007), Kolmogorov Complexity (Li and Vitányi, 2009), Algorithmic Probability (Solomonoff, 1997), as well as in Algorithmic Complexity (Chaitin, 2003, 1975). Kolmogorov, in particular, sought to define computational complexity in terms of the smallest (i.e., the most elegant) string of instructions in some language that can be used to reconstruct another string. If the language is FOL, and the given string is a list of formulas, Kolmogorov’s goal would then be to try to find the most elegant set of axioms from which we can derive those formulas. The sad fact is that finding the smallest such set of axioms is an undecidable question. There is no way to know if the current set of axioms is the most elegant by any mechanical method. It is also known that the shortest length of a formula is language dependent; as alluded to above, some formal languages may be better for describing certain concepts than others.

\footnote{We have that \((A \lor B) \leftrightarrow (A \land \neg B) \lor (\neg A \land B) \lor (A \land B), A \leftrightarrow (A \land \neg B) \lor (A \land B),\) and \(B \leftrightarrow (\neg A \land B) \lor (A \land B).\) This, in turn, gives us \(P(A \lor B) = P(A \land \neg B) + P(\neg A \land B) + P(A \land B), P(A) = P(A \land \neg B) + P(A \land B), P(B) = P(\neg A \land B) + P(A \land B),\) respectively. Combining these three terms and rearranging gives the theorem.}
Thus although the idea of using elegance to objectively assign probabilities to sentences is intriguing, it is beyond the scope of this current work to evaluate. I leave it as a item for future study, and I see it as the frontier of probability theory.

3.2.8 A Paradox?

Probability theory speaks of numbers (probabilities) and functions (the probability assignment operator $P$); in order to use those terms we must already have accepted ZFC as a deductive system, since ZFC derives those concepts as theorems. Yet, if our framework allows all formal deductive systems to be viewed as probabilistic logics, we have a potential problem if we place a probability of less than 1 on any of the ZFC axioms. This would cascade a probability of less than 1 down to any theorem implied by it. For the sake of argument, let’s assume that an axiom (say, the Axiom Schema of Restricted Comprehension) was given a probability of 0.8 which then implies that there is a probability 0.8 that such a thing as a function exists. We now try to assign probabilities with $P$ to a statement (say, again, to the Axiom of Schema of Restricted Comprehension that we started with), but now we have that the probability of the axiom having a probability of 0.8 is itself 0.8! Our deductive tree is no longer a tree, but a loop. This self-referential iteration would then continue forever, and it would seem to imply that, due to the effect of multiplying probabilities of less than one, that our final probability for both the axiom and its theorems would be 0.

How does one handle this? Does it even make any sense? The only way out of this that I can see is the obvious one: we are simply not allowed to assign probabilities other than 0 or 1 to the axioms of ZFC. I am not sure if this is a peculiarity of ZFC, or if perhaps the same problem and solution would exist for other axiomatic systems (say, perhaps the logical axioms of FOL themselves....?). I have not analyzed this issue any further, but it is interesting question for future research.
3.3 Probabilistic Deduction and the Problem of the Catch-all Hypothesis

With the concept of probability we can add more structure to a logical language by looking at the probability of deduced theorems. This can be seen in the following two theorems related to conditional probability which we derive from the above axioms:

Theorem 4: **Total Probability Rule**

\[ P(T) = P(H)P(T|H) + P(\neg H)P(T|\neg H) \]

Theorem 5: **Total Probability Rule (Multiple Mutually Exclusive Hypotheses)**

\[ P(T) = \sum_j P(H_j)P(T|H_j) \]

These theorems tell us how to calculate the probability of a theorem in a probabilistic deductive system given the probability of each hypothesis as well as the conditional probability of the theorem given that the hypothesis is true; if we have all the required probabilities, we can calculate the value of \( P(T) \). The problem comes, however, with what is known as the ‘catch-all’ hypothesis. This is the \( \neg H \) term in Theorem 4 and represents every possible hypothesis other than \( H \). In some applications we know exactly what the possible options are; e.g., the hypothesis that it will rain tomorrow and its negation (that it won’t rain), or else the hypothesis that a coin will come up heads and its negation (that it will come up tails). In other cases \( \neg H \) includes every possible hypothesis that might explain observations other than the ones which have been thought of so far. Such is the case in the philosophy of science where we might wish to assign probabilities to a scientific hypothesis. Salmon states that this would amount to assigning probabilities to all other scientific theories, including those that

---

\(^6\)To derive this theorem we assume that \( H_1, H_2, \ldots, H_M \) are mutually exclusive hypotheses, then take the following tautology \( T \leftrightarrow T \land (H_1 \lor H_2 \lor \cdots \lor H_M) \), distribute through with the conjunction, and apply theorems 2 and 3 to get \( P(T) = P(T \land H_1) + P(T \land H_2) + \cdots + P(T \land H_M) \). From above we have that \( P(T) = P(H_1)P(T|H_1) + P(H_2)P(T|H_2) + \cdots + P(H_M)P(T|H_M) \). This derivation obviously applies to the previous theorem, also.
haven’t been thought of yet (Curd et al., 2013, pg. 535); which seems to imply having knowledge of the future of science. If, however, we are able to assign probabilities to all possible formulas based on their syntax, as we considered above in Section 3.2.7, then this might actually be possible. Since we are using a formal language, and the number of possible well-formed formulas in the language is countable, the question becomes whether or not we can distribute probabilities over a countable set in such a way that the sum remains finite. If this were feasible, then by first assigning probabilities to the hypotheses being actively considered based on their syntax, one could then sum their probabilities and subtract from 1, thereby calculating the probability of the catch-all. Regarding the total probability not summing to 1, an approximation that would make this feasible would be to simply truncate the list of possible expressions in the language beyond some subjectively pre-determined maximum string length. This is ad-hoc, but might prove itself effective for approximating $P(\neg H)$.

Although theoretically interesting, it is actually a moot point whether or not $P(\neg H)$ can be calculated, since this is only half the problem. In order to calculate $P(T)$ we would still need to assign a probability to $P(T|\neg H)$, which is even more intractable. Since we have not used an inductive process to create the hypotheses contained in the catch-all, we have no idea whether an observation is implied by $\neg H$, or not. In the example of Ptolemaic versus Copernican astronomy, this would be like asking the question whether a yet-to-be-thought-of hypothesis predicts parallax of the stars. We cannot say what the conditional probability of theorems of a hypothesis would be without knowing what the hypothesis is. My conclusion is that this problem is utterly intractable and fundamentally limits our ability to analyze the probability of theorems, even if a probability could be assigned to $P(\neg H)$. Since there appears to be no solution, a simple and straight-forward alternative is to ignore the catch-all hypothesis and assume we only have the hypotheses we have formulated. In this case
we can remove the catch-all term, distribute the total probability over the remaining hypotheses, plug in values for the other probabilities in Theorem 5, and calculate the value of $P(T)$. Technically, this will not give us a value of probability (since it is incomplete); it will, however, effectively rank the theorems and show us which ones are more likely than the others. Thus we see that except for the special case where we actually have probabilities legitimately assigned to all hypotheses, and related conditional probabilities assigned to theorems given that hypotheses are true, the absolute probability of a theorem is not a well-defined concept.

### 3.4 Examples

We will now show what the deductive systems from Section 2.3 look like when probabilities are assigned to both premises, and to theorems conditioned on premises. We first present the assignment of probabilities in tables, and then express them graphically via the thicknesses of bounding boxes and arrows in Figures 3.1 through 3.4. In contrast to the Figures in Section 2.3, here the thickness of an arrow is proportional to the conditional probability of a theorem given a premise, and the thickness of a bounding box is proportional the probability of a theorem or hypothesis. $P(T)$ will be greater when either the probability of a hypothesis from which it is deduced, or the conditional probability that connects the hypothesis and theorem, or both, is large. I assign the probabilities arbitrarily; for the purposes of these examples it is not relevant how this was done. The goal, rather, is to show how probabilistic logic can handle formal inference in a variety of applications including philosophy of science, law, medicine, and statistics.

The first example returns to the case of competing theories of astronomy, and is shown in Figure 3.1, with the corresponding values in Table 3.2. There are three possible hypotheses: Ptolemaic, Copernican, as well as the catch-all. We ignore the
catch-all hypothesis, and begin with the Ptolemaic theory \((H_1)\). It predicts that there will probably not be any observed parallax \((\neg T_1)\), it also predicts the motion of planets with high likelihood \((T_2)\), and it asserts that there should be little variation in brightness of planets throughout the year \((\neg T_3)\), again with high likelihood. The Copernican hypothesis \((H_2)\), on the other hand, predicts \(T_1\), \(T_2\), and \(T_3\), all with high likelihood, although \(T_1\) with slightly lower likelihood since it maintained the possibility that stars were so far away that parallax would not be detectible. The fact that both hypotheses have small conditional probabilities to observe the opposite of what they should predict is due observational error, e.g., the naked eye of a Ptolemaic astronomer could mistakenly observe parallax when it really isn’t there. Ptolemaic and Copernican hypotheses are assumed to have the same a priori probability. This is shown by giving them both moderately thick bounding boxes, signifying a value of 0.5 each. Based on the probabilities, the values in the bottom part of Table 3.2 are derived using Theorem 5. These numbers show the relative degrees of probability of the various theorems and their negations. \(T_2\) has the highest values (99%) because both theories are capable of predicting planetary motion to a high degree. The second most probable theorem is that there is no observed parallax \((\neg T_1)\). This is predicted by Ptolemaic astronomy, but is slightly unlikely under the Copernican view. The third most probable theorem is the lack of varying brightness of planets during their motion in the sky. It is slightly in favor of \(\neg T_3\) over \(T_3\) because the Ptolemaic view is slightly stronger in its certainty than the Copernican one. (Again, this is just for illustrative purposes; in reality these probabilities may have been very different.)

The next example returns to the medical diagnosis of the cause of Parkinson-like symptoms. Figure 3·2 shows the same information as Figure 2·4, but with the additional information in the form of the thickness of the bounding boxes and arrows. The causal hypotheses are: \(H_1\), the Parkinson gene, \(H_2\), Lewy bodies in neurons in the
Substantia Nigra, and $H_3$, chemicals in drugs used by addicts. $T_1$ says that symptoms appearing at a young age, $T_2$ is the sudden onset of symptoms. In this case we have that some hypotheses have greater prior probability that others. In particular, having a Parkinson gene is rarer than the normal development of Lewy bodies in the neurons of the brain. The case of chemicals in hard drugs is an even rarer occurrence than the Parkinson gene. This information is reflected in the sizes of the bounding boxes around the three hypotheses. The conditional probabilities also differ; as mentioned in Section 2.3, the ‘chemicals in drugs’ explanation only has the characteristic of ‘sudden onset’ in $T_2$. Thus this theorem has a conditional probability of 1, and the conditional probability of its negation is 0 (and therefore no arrow is drawn to that possible outcome). A similar situation holds for the ‘Lewy body’ hypothesis and the prediction of symptoms appearing at a young age. The most probable theorems in this case are having symptoms appear with slow onset in older patients. This makes intuitive sense, since the most likely cause a priori is that there are Lewy bodies in the neurons of the Substantia Nigra, which usually implies these theorems with high likelihood.

The third example comes from the situation where Bob is potentially cheating on his wife. Again, this example relates to determining the guilt or innocence of an accused party, and thus has applications to law. We show graphically in Figure 3·3 the information from Table 3.4. The causal hypotheses are: $H_1$, Bob is cheating on her, or $H_2$, a constellation of other hypotheses exonerate Bob from wrong-doing. The theorems are: $T_1$, Bob was seen at a bar with another woman, $T_2$, Bob had lipstick on his collar when he got home from work, and $T_3$, that Bob had a large un-explained bill with a jeweler. The observations are equally possible, but very unlikely, under either of the hypotheses, hence they have low conditional probabilities. This can intuitively be understood as follows: If Bob is cheating, he will probably be covering his tracks,
and thus it is unlikely he will be seen in a bar with his mistress, or come home with lipstick on his collar, or have a large bill with a jeweler. If he is not cheating, the probability is equally unlikely any of those things will occur for obvious reasons. Thus in this case the most probable theorems are the negations $\neg T_1$, $\neg T_2$, and $\neg T_3$. The only visibly interesting feature in the figure is the fact that Bob’s alternate hypothesis (of not cheating) is much more complicated than the simple ‘cheating’ one. In fact, since it contains more sub-hypotheses, it is less parsimonious (Section 4.1.1, below), and will thus have lower prior probability. Since the observations will do nothing to determine Bob’s guilt or innocence because of their equal likelihoods, we can see that Bob will be determined to be guilty solely on the fact that his explanation for the observations is more convoluted than the simple hypothesis of him cheating (see Section 4.3, below).

The final example returns to the case of polynomial fitting. We have limited ourselves to only two possibilities for each polynomial for both theoretical reasons, and for clear exposition\(^7\). We note that the probability of $L_1$ is higher than $L_2$, since it is syntactically simpler (more elegant). I arbitrarily set the conditional probability of the coefficients given the choice any polynomial to the value 0.5. The possible coefficients have been carefully chosen to match the theorems. In Figure 2-6 an implication arrow only came from a set of coefficients to one or the other of the choices in each theorem (either $T$ or $\neg T$). We relax this constraint now, since we can assign probabilities to any $x - y$ pair from any set of coefficients; the further away the $x - y$ pair is from the normally predicted value, the lower its conditional probability.

---

\(^7\)In real applications there may be a continuous distribution of possible coefficient values. Since every example presented so far has only had a finite number of hypotheses and a finite number of implications arising from a hypothesis, accounting for a countable number would move us into a more complicated regime. We would have to examine whether all the axioms can be extended to infinite collections sentences. Since that would be a daunting task and out of the scope of this work, I take the prudent path, and limit this case of polynomial fitting to a very small number of hypotheses and theorems.
Thus we have more arrows in Figure 3·4 than in Figure 2·6. In Table 3.5 we have assigned much higher conditional probabilities to the negations of theorems. This is again due to the fact that we are allowing the possibility of many $x - y$ observations to be consistent with a single set of coefficients. Thus there may be more total probability distributed to the disjunction of other possible $x - y$ pairs than to the most likely one predicted by a pair coefficients for a given polynomial. This, in turn, makes the negations of the theorems more probable than the theorems themselves in the case of polynomial fitting.
Table 3.2: Top table: probabilities of scientific hypotheses and conditional probabilities for observed astronomical phenomena given the hypotheses. Bottom table: probabilities of theorems calculated using Theorem 5 with the values above.

| $P(H)$ | $P(T|H)$ | $T_1$ | $\neg T_1$ | $T_2$ | $\neg T_2$ | $T_3$ | $\neg T_3$ |
|--------|----------|-------|-------------|-------|-------------|-------|-------------|
| 0.5    | $H_1$    | 0.01  | 0.99        | 0.99  | 0.01        | 0.01  | 0.99        |
| 0.5    | $H_2$    | 0.6   | 0.4         | 0.99  | 0.01        | 0.95  | 0.05        |

<table>
<thead>
<tr>
<th>$P(T)$</th>
<th>$T_1$</th>
<th>$\neg T_1$</th>
<th>$T_2$</th>
<th>$\neg T_2$</th>
<th>$T_3$</th>
<th>$\neg T_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.305</td>
<td>0.695</td>
<td>0.99</td>
<td>0.01</td>
<td>0.48</td>
<td>0.52</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3-1: A graphical representation of the information in Table 3.2.
Table 3.3: Top table: probabilities of hypotheses and conditional probabilities given the hypotheses related to Parkinson-like symptoms. Bottom table: probabilities of theorems calculated using Theorem 5 with the values above.

| $P(H)$ | $P(T | H)$ | $T_1$ | $\neg T_1$ | $T_2$ | $\neg T_2$ |
|--------|------------|-------|-------------|-------|-------------|
| 0.05   | $H_1$      | 0.8   | 0.2         | 0.1   | 0.9         |
| 0.94   | $H_2$      | 0.1   | 0.9         | 0.01  | 0.99        |
| 0.01   | $H_3$      | 0.5   | 0.5         | 1.0   | 0.0         |

<table>
<thead>
<tr>
<th>$P(T)$</th>
<th>$T_1$</th>
<th>$\neg T_1$</th>
<th>$T_2$</th>
<th>$\neg T_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.139</td>
<td>0.861</td>
<td>0.244</td>
<td>0.9756</td>
</tr>
</tbody>
</table>

Figure 3.2: A graphical representation of the information in Table 3.3.
Table 3.4: Top table: probabilities of hypotheses and conditional probabilities given the hypotheses related to the possibility of Bob cheating on his wife. Bottom table: probabilities of theorems calculated using Theorem 5 with the values above.

| $P(H)$ | $P(T|H)$ | $T_1$ | $\neg T_1$ | $T_2$ | $\neg T_2$ | $T_3$ | $\neg T_3$ |
|--------|----------|-------|-------------|-------|-------------|-------|-------------|
| 0.75   | $H_1$    | 0.2   | 0.8         | 0.1   | 0.9         | 0.01  | 0.99        |
| 0.25   | $H_2$    | 0.2   | 0.8         | 0.1   | 0.9         | 0.01  | 0.99        |

<table>
<thead>
<tr>
<th>$P(T)$</th>
<th>$T_1$</th>
<th>$\neg T_1$</th>
<th>$T_2$</th>
<th>$\neg T_2$</th>
<th>$T_3$</th>
<th>$\neg T_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.8</td>
<td>0.1</td>
<td>0.9</td>
<td>0.01</td>
<td>0.99</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.3: A graphical representation of the information in Table 3.4.
Table 3.5: In this case we must split our hypotheses into postulated laws and the coefficients each law implies. Top table: prior probabilities on the laws and the conditional probabilities of the coefficients related to polynomials. Middle table: prior probabilities on the conjunctions of the laws and their coefficients, along with the likelihoods of observations of empirical data given each hypothesis. Bottom table: probabilities of theorems calculated using Theorem 5 with the values above.
Figure 3-4: A graphical representation of the information in Table 3.5.
Chapter 4

Abduction

Abduction is the third category of inference; it is inference to the optimal explanation of a potentially underdetermined system, sometimes referred to as “reasoning backwards”. It is the kind of inference usually associated with statisticians and inverse theorists in the applied sciences, who, rather than deducing theorems or discovering laws of nature, seek to devise methods for choosing an optimal hypothesis that explains the data from a set of possible choices. It has the most uncertain status among the three categories of inference. It resembles deduction in the sense that, given enough observations and constraints in the system, one can deduce a unique optimal hypothesis. It resembles induction in that there must be an act of inspiration in order to create a set of possible hypotheses to choose from and assign probabilities, and thereby constrain the system enough to allow deduction.

Using the framework of probabilistic logic I will show that the optimal abductive hypothesis falls out naturally in the form of Bayes rule. I will examine a couple of common criticisms of the Bayesian paradigm — specifically I will return to the problem of the ‘catch-all’ hypothesis in the abductive setting, and then look at the Problem of Auxiliary Hypotheses. I next apply Bayes theorem to the same examples from Chapters 2 and 3, and examine the results.
4.1 Bayesian inference

In the framework of probabilistic logic, the abductive solution is an almost trivial deduction from the axioms and theorems of Chapter 3. Using the axiom of conditional probability again we have:

\[ P(T \land H) = P(T)P(H|T) \] (4.1)

Combining this with Equation 3.5 gives us Bayes rule:

\[ P(H|T) = \frac{P(T|H)P(H)}{P(T)} \] (4.2)

We may then use Theorem 5 for the case of multiple hypotheses \( H_j \) to arrive at

**Theorem 6: Bayes Theorem**

\[ P(H_1|T) = \frac{P(T|H_1)P(H_1)}{\sum_j P(T|H_j)P(H_j)} \]

\( P(H_1) \) is then known as the *prior* probability of a particular hypothesis \( H_1 \); it captures the state of knowledge about \( H_1 \) before \( T \) is observed. \( P(H_1|T) \) is known as the *posterior* probability of \( H_1 \) and measures our augmented state of knowledge about \( H_1 \) given that \( T \) was observed. \( P(T|H_1) \) is known as the *likelihood function* of \( H_1 \) on \( T \), and expresses the probability of the observation given that a hypothesis is true.\(^1\)

\( P(T) \) is called the *expectedness* (Curd et al., 2013, pg. 531) of the observation. If an unlikely event occurs (low \( P(T) \)) the posterior probability will be high; this is in line with our intuition — an unlikely event will be very informative about the system.

The abductive inference, then, is to pick the hypothesis, \( H_j \), that has the highest posterior probability.\(^2\) A problem with doing this, already mentioned above, was the

---

\(^1\)For a fixed hypothesis \( H_j \), \( P(T|H_j) \) is a conditional probability defined on \( T \). If, however, we fix \( T \) and examine the values of \( P(T|H_j) \) for different choices of \( H_j \) it is not a probability, hence the term *likelihood*.

\(^2\)It is often the case in applications of abduction in inverse theory and statistics that we don’t want
difficulty in calculating the probability of $P(T)$ from Equation 5, which may occur for either practical reasons (too many terms to deal with) or theoretical reasons (the intractable catch-all hypothesis). For the case of theoretical intractability, the unsatisfying solution suggested in Section 3.3 was to simply ignore the catch-all and pretend that the hypotheses being considered are the only ones that exist. Probability is then distributed over them and $P(T)$ can be calculated. This would not lead to legitimate posterior probabilities, but would effectively allow ranking of the hypotheses. Salmon (2013, page 536), however, uses the alternate procedure of calculating ratios of posterior probabilities for the different hypotheses being considered as a way to avoid this problem. For competing hypotheses $H_1$ and $H_2$, he has:

$$\frac{P(H_1|T)}{P(H_2|T)} = \frac{\frac{P(T|H_1)P(H_1)}{P(T)}}{\frac{P(T|H_2)P(H_2)}{P(T)}} = \frac{P(T|H_1)P(H_1)}{P(T|H_2)P(H_2)}$$

(4.3)

If the ratio is greater than unity, the first hypothesis is deemed more explanatory of the observations, and vice versa if less than unity.

If the problem of calculating $P(T)$ comes from the practical consideration of having too many hypotheses to deal with, I agree with Salmon’s approach. In that case we know that $P(T)$ has a numerical value — we just can’t calculate what it is. Thus we can save ourselves trouble and still find the best choice of hypothesis without worrying about the term. If, however, we are unable to calculate it for theoretical reasons (i.e., the catch-all hypothesis from Section 3.3), Salmon’s approach may not be possible. It was shown earlier that the term $P(\neg H)P(T|\neg H)$ was doubly intractable. I put forth an idea that might allow for practical calculation of $P(\neg H)$ if objective assignment

---

the hypothesis with highest probability. Such a case happens when the hypotheses have some sort of natural ordering and distance metric defined on them (say, the real numbers). The probabilities then take the form of a distribution over the real line. The best abduction might then be to take some sort of moment with respect to the distribution, such as the mean or the median, rather than the mode (maximum value). Which one to choose depends on the application, and the risk associated with choosing the wrong hypothesis. For instance, when money is at stake, one might be better choosing the mean so that on-average losses will be minimized.
of probability based on elegance principles were possible (Section 3.2.7). It seemed, however, that even that concept would be of no use in calculating $P(T|\neg H)$. The question then becomes: “Does the value of $P(T|\neg H)$ actually exist, even if we could never access it?”. If not, then $\frac{P(T)}{P(T)}$ is undefined, and the above equality is nonsense. A counter-argument, however, might be that with infinite time, for any sub-hypothesis in the catch-all that we might postulate, a likelihood term could be constructed for it. In this case Salmon’s idea works. I do not know how to analyze this metaphysical question any further. Since assuming the existence of a value for $P(T|\neg H)$ leads to useful algorithms, in what follows I’ll assume it does exist, and thus the ratio of probabilities involving a catch-all hypothesis still has meaning.

4.1.1 Quantitative Parsimony and the Problem of Auxiliary Hypotheses

We examined one form of simplicity in Section 3.2.7, above. There elegance was a measure of how short a sentence was in a language, and allowed the possibility of an objective assignment of probability to all sentences in the probabilistic logic. Here we examine a different kind of simplicity known as parsimony. The principle has existed since ancient times, and is most commonly known by the name Ockham’s Razor, in honor of William of Ockham, an English Franciscan friar (Baker, 2013). Ockham originally argued that nothing should be posited as being true unless it is a self-evident truth, it is learned through experience, or has been revealed by God. Modern interpretations of his view have it as saying that of all explanations that fit a set of observations equally well, we should always pick the one that postulates the fewest number of entities. The definition of what qualifies as an ‘entity’ has led philosophers to further break the concept into two sub-categories: quantitative and qualitative parsimony (Baker, 2013). Quantitative parsimony refers to postulating fewer objects
in the universe of discourse; qualitative parsimony refers to the number of kinds of objects being postulated.

Quantitative parsimony can be clearly expressed in a probabilistic logic. If a hypothesis \( H \) can be split up into the conjunction of numerous independent hypotheses, \( H \leftrightarrow H_1 \land H_2 \land \cdots \land H_M \), we can assign probabilities to obtain

\[
P(H) = P(H_M | H_1 \land \cdots \land H_{M-1}) P(H_{M-1} | H_1 \land \cdots \land H_{M-2}) \cdots P(H_3 | H_1 \land H_2) P(H_2 | H_1) P(H_1)
\]

(4.4)

Since each conditional probability must be less than or equal to unity, the product will have a strong tendency to become very small as we increase the number of sentences that make up our hypothesis. Therefore we can see that maximizing parsimony is equivalent to minimizing the number of independent hypotheses used to explain observations. We can add more detail to this concept through our hypothesis involving predictive laws in FOL:

\[
H \leftrightarrow (\exists c \forall x \exists y Lcxy) \land (\exists c(c = c'))
\]

We recall that the boldface notation on \( c \) meant an array of objects, so \( (\exists c(c = c')) \) can be expanded as \( (\exists c_1(c_1 = c_1')) \land (\exists c_2(c_2 = c_2')) \land \cdots \land (\exists c_M(c_M = c_M')) \). Thus this hypothesis can be rewritten as

\[
H \leftrightarrow \forall x \exists y Lcxy \land (\exists c_1(c_1 = c_1')) \land (\exists c_2(c_2 = c_2')) \land \cdots \land (\exists c_M(c_M = c_M'))
\]

(4.5)

Since, given the law, each postulated coefficient can be considered a separate hypothesis, as we increase the number of coefficients in the law, it will become less parsimony.

---

\(^3\)Although the concept of qualitative parsimony is more prominent in the philosophical literature, I currently don’t know how to incorporate it into the probabilistic logical framework, and therefore will not address it here.
monious and the probability of $H$ will decrease. This holds regardless of whether the individual coefficients ($\exists c_1(c_1 = c'_1)$), ($\exists c_2(c_2 = c'_2)$), etc., are dependent or independent of each other. It also holds regardless of how we assign probabilities; adding more parameters will tend to lower the probability of $P(H)$ because we will be multiplying more terms together.

Although increasing the number of hypotheses that imply the observations leads to a lowering of the posterior probability, this assumes that there is no associated change in the likelihood and expectedness values. If, however, we choose to include an extra hypothesis that, in conjunction with the current hypotheses, either greatly increases the likelihood or lowers the expectedness, we may actually increase the posterior probability. This is known as the Problem of Auxiliary Hypotheses. It is a common way out of any argument where the observations seem to imply one’s belief is wrong. By simply including an appropriately chosen new hypothesis, no matter how outlandish, one can increase their posterior probability. A good example of this is the case of Bob, above, who needs to invent several hypotheses to reach the same level of posterior probability as the hypothesis that he is cheating. Perhaps if he had been more clever, he could have added even more hypotheses that would have made his alibi perfect and increased his posterior probability above the cheating hypothesis. A different example would be the Ptolemaic method of epicycles for accounting for the planetary motions; to achieve higher and higher accuracy, the method has to postulate more and more epicycles upon epicycles in order to explain observations.

The technique to counter this problem is two-fold: first, the new hypotheses should be chosen such that they are falsifiable by a new observation that can be performed. I.e., they should have an high-likelihood implication to a theorem that can be tested. If the new auxiliary hypothesis is untestable by definition, it should not be admitted. The second method is an asymptotic one; by collecting more observations the likeli-
hoods will eventually dominate the priors, thus nullifying the benefit of any kind of auxiliary hypothesis. A proof of this can be sketched by expanding the denominator in Bayes rule and comparing it with the numerator. For some hypothesis $H_j$ we have:

$$P(H_j|T) = \frac{P(T|H_j)P(H_j)}{P(T|H_1)P(H_1) + P(T|H_2)P(H_2) + \cdots + P(T|H_M)P(H_M)} \quad (4.6)$$

where $T = T_1 \land T_2 \land \cdots \land T_N$. Since the likelihood terms $P(T|H_j)$ for any hypothesis $H_j$ can be expanded as a product of $N$ separate terms (one for each observation), we can see that its value will decrease with $N$. As an example, I will assume that hypothesis $H_1$ tends to be a better explanation for observations, so the rate at which its likelihood term decreases is slower than for other hypotheses. I will now display the asymptotic effect of differing rates of convergence for different hypotheses in a graphical way, by repeating Equation 4.6 for hypothesis $H_1$, but with a smaller font for the terms that tend towards zero faster.

$$P(H_1|T) = \frac{P(T|H_1)P(H_1)}{P(T|H_1)P(H_1) + P(T|H_2)P(H_2) + \cdots + P(T|H_M)P(H_M)} \quad (4.7)$$

We can see that the value of $P(H_1|T)$ will tend towards 1. If we examine another hypothesis, $H_2$, that tends to explain observations worse, we have the following visual:

$$P(H_2|T) = \frac{P(T|H_2)P(H_2)}{P(T|H_1)P(H_1) + P(T|H_2)P(H_2) + \cdots + P(T|H_M)P(H_M)} \quad (4.8)$$

Here we see that the value of $P(H_2|T)$ will tend towards zero. If we instead formed a ratio of the posteriors for $H_1$ to $H_2$, we would see that it would tend towards infinity with increasing observations.

### 4.2 The Error-statistics Approach

A problem with criticizing the Bayesian viewpoint is that there are few alternatives that lead to a formal calculus for abductive inference. One alternative is to just give
up and say it proceeds in an ad-hoc evolutionary fashion, where by random chance better hypotheses tend to survive longer than worse ones. This view was advocated by Thomas Kuhn (Kuhn, 2012) in his post-script when discussing the progress of science. I reject Kuhn’s view; it is my opinion that if a philosopher really understands something about how science progresses he or she should be able to write it out in a formal logical language — after all, formal languages exist to capture our informal reasoning in a way that allows it to be checked for accuracy and then applied to new situations. Although he denies being irrational, Kuhn’s point of view essentially says there is no rational order to be found that will codify how human reasoning is performed, and, as a corollary, no way to improve upon it.

One possible alternative for formalizing scientific progress is the “Error-statistics” approach of Deborah Mayo (Mayo, 2013). This approach is closely associated with the classical hypothesis testing methods put forward by the statisticians Ronald Fisher, Jerzy Neyman, and Egon Pearson in the early twentieth century. It is not possible to give an adequate review of their methods here. The points relevant to our discussion are the following: Fisher’s approach used an idea similar to the FRPS, above. It was an absolute test of whether or not the result of an experiment is likely under an assumed “null hypothesis”\textsuperscript{4}. If it is not, the null hypothesis must be rejected. Neyman and Pearson used a variant of this principle where hypotheses are only compared to each other in the form of a ratio. In relation to the ratio test of Bayes theorem, these methods could be seen as limiting cases where prior probabilities become uniform and cancel out, thereby leaving only a comparison of likelihood terms.

Mayo argues that since this method removes the possible subjective bias that comes from the prior probabilities, it is better. She makes the point that in order to get our priors in the first place, we usually take a frequentist approach and use empirical data about hypotheses from the past (or at least analogs, if there is no

\textsuperscript{4}The term “null hypothesis” usually refers to a hypothesis that one wishes to disprove.
empirical data), and by some non-Bayesian method we calculate the prior probabilities for each hypothesis (Mayo, 2013, page 555). She is correct in stating this, but fails to realize that the same problem applies to either the Bayesian or Error-statistics approach when calculating likelihood functions. They are also based on empirical data, and we would use the same method calculate their values.

As a separate criticism of the Bayesian approach, Mayo also brings up the idea of a utility function as part of the process of choosing a best theory (Mayo, 2013, page 552). This would be a function that would take the posterior values over the hypotheses and weight them according to the subjective values of the community examining it. These values may be personal prejudices, political views, etc. Mayo argues that although subjectivity can be modeled by the Bayesian approach in the form of the priors, it is not possible to compare expected utilities. This is a more legitimate criticism, but it again fails for the reason that it can be turned against the Error-statistics approach, also. Instead of weighting posterior probabilities by the utility function, one could weight the likelihood functions. This would render the Error-statistics ratio test equally intractable.

Although likelihood ratios are the most common way to compare hypotheses in the Error-statistics approach, the absolute values would work just as well. In Tables 4.1 through 4.5, below, these values are represented in the fourth column. For some of the examples the most likely hypothesis (i.e., the one chosen by the Error-statistics approach) is the same as the one chosen by the Bayesian method. Notable exceptions are Tables 4.2 and 4.3 for the case of medical diagnosis, where the abduced hypothesis is different, and Table 4.4, where the Error-statistics approach is indeterminate.

It might seem that the Error-statistics approach is superior because it removes the bias of prior probabilities, but this depends on whether one sees this as a plus or a minus. Since prior probabilities were shown above to capture the theoretical virtue
simplicity (via both elegance and quantitative parsimony), it seems that we are losing something if we abandon the Bayesian approach in favor of this alternative method.

4.3 Examples

In order to gain further insight into the Bayesian method it is instructive to take a few clear examples containing several hypothesis along with observations, assign them reasonable probabilities, and see how the two criteria differ on their decisions for the optimal hypothesis. This is done in the following section where we return to the examples from earlier chapters.

In Tables 4.1 through 4.5 we show probabilities related to abduction via Bayes theorem. The second column of each table is the posterior probability of a hypothesis given the observations. In the third column are the prior probabilities of each hypothesis. The fourth column gives the joint likelihood of observations (here we are assuming they are independent of each other and can be calculated as the product of the individual likelihoods of each observation given the hypothesis). This column corresponds to the Error-statistics inference criterion for ranking hypotheses. The fifth column gives the expectedness of all observations co-occurring. The posterior probability is obtained by multiplying the prior by the likelihood term, and then dividing by the expectedness. It should be noted that one cannot calculate the values of expectedness by assuming that $P(T_1 \land T_2 \land T_3) = P(T_1)P(T_2)P(T_3)$ and multiplying together the values of the $P(T_j)$ terms on the right hand side of the top of Tables 3.2 through 3.5. That would be incorrect; the observations are conditionally independent of each other given a hypothesis, but not unconditionally independent of each other. This makes intuitive sense, since the probability of the $T$’s may have strange dependencies among themselves, e.g., with the medical diagnosis example below, slow onset of symptoms tends to happen when symptoms appear at old age, while rapid onset
tends to happen when they appear at a younger age. The correct way to calculate $P(T_1 \land T_2 \land T_3)$ is therefore to use Theorem 5.

Figures 4·1 through 4·5 show the same examples from Figure 3·1 through 3·4, except we have now put a red asterisk by the theorems that were actually observed, and we have ‘pruned the deductive tree’ to remove any implications that don’t lead to these observations. We have also listed the posterior probability in red letters next to each of the hypotheses.

Beginning with Table 4.1, we see the probabilities associated with abducting either Ptolemaic or Copernican astronomy as being the best hypothesis based on astronomical data. Here we have assumed a priori that both theories were on equal footing, and thus assigned a probability of 0.5 to each. As an example of calculating the expectedness term from Theorem 5, we plug in prior and likelihood values to the following equation

$$P(\neg T_1 \land T_2 \land T_3) = P(H_1)P(\neg T_1|H_1)P(T_2|H_1)P(T_3|H_1) + P(H_2)P(\neg T_1|H_2)P(T_2|H_2)P(T_3|H_2)$$

and show the result in the last column of Table 4.1. Then, using Bayes theorem we obtain the posterior values in the second column. From these we can see from a Bayesian viewpoint that Copernican astronomy is greatly favored over Ptolemaic — so much so that Ptolemaic astronomy has almost a negligible chance of being true. In this particular example of Ptolemy versus Copernicus, since both the prior and expectedness terms remain constant for both hypotheses, we can see that the ratio of posterior probabilities will only differ from the ratio of likelihoods by a constant factor. Therefore the Error-statistics criterion gives us the same answer as Bayesian inference (although Bayes can give actual probabilities, while Error-statistics can only give likelihoods).
Although this analysis of competing scientific theories is interesting and may accurately describe the process of choosing a superior theory in principle, a question to ask is whether it can be used in practice. I know of no case where it actually has been done, but I see no reason that it could not be attempted. The fact that there is no universally agreed upon method to assign probabilities does not, and should not, preclude it from being used. If two scientific communities can agree upon the same values for the assignment of probabilities, it would be trivial to decide on the superior theory; but even if they don’t agree (and thereby arrive at different inferences) they can at least retrace their steps and examine why their posterior probabilities on the hypotheses are different. This would perhaps point out over-confident prior probabilities that are not founded in any previous experience or logical argument.

And, furthermore, even if both sides of an argument happen to agree on the prior probability values of their respective hypotheses that are input into a probabilistic inference method, the losing side would not have to give up its normal science. If, over time, it could be seen that adding observations was shifting the ratio of probabilities in their favor, it would be logical to continue investigation until the maximum posterior probability switched to their hypothesis. This is reminiscent of the Kuhnian idea that scientists working on a new hypothesis should be allowed to continue work, even if early results show that their system does not fit the observations as well as the entrenched paradigm does. Indeed, Kuhn in his later years agreed, at least in concept, with the Bayesian viewpoint as a method for deciding between scientific theories (Kuhn, 2013, page 101):

...each scientist chooses between competing theories by deploying some Bayesian algorithm which permits him to compute a value for \( P(H|T) \), i.e., for the probability of a hypothesis \( H \) on the observation \( T \) available both to him and to other members of his professional group at a particular
period of time. [My terms and notation.]

We now turn to the second example of medical diagnosis for Parkinson-like symptoms, where we will examine two separate cases: first, the observed symptoms are in an older patient who had a sudden onset; second, we have a young patient with slow onset of symptoms. Looking at Table 4.2 for the first case we see that the posterior probability on the hypotheses is greatest for \( H_2 \) (that there are Lewy bodies in the neurons of the Substantia Nigra — the classic cause of Parkinson's disease). The posterior probabilities for the second case are shown in Table 4.3, and also makes the same abductive choice of hypothesis given the observations. Since prior probabilities are different, the Error-statistics viewpoint gives quite different answers for this example. For the first case it assigns the highest likelihood to \( H_3 \), the hypothesis that chemicals in hard drugs caused the symptoms. For the second case it assigns the highest likelihood to \( H_1 \), the hypothesis that the Parkinson gene is the cause. This example therefore shows where Bayesian and Error-statistics viewpoints can give very different answers. The difference comes from the prior probabilities which bias the abductive choice of optimal hypothesis.

Next we turn to the example of Bob, who may or may not be cheating on his wife. Here the observations are that Bob was seen with another woman at a bar \( (T_1) \), he came home from work with lipstick on his collar \( (T_2) \), and he had a large un-expected bill from a jeweler \( (T_3) \). There are two possible explanations (hypotheses) for these facts; the first is much simpler than the second. Since it seems reasonable to assign higher probability to simpler hypotheses, we give the straightforward explanation that Bob is cheating a probability of 0.75 that it is true. Bob’s complex explanation, since it must postulate many auxiliary hypotheses, is given a lower probability of 0.25. The likelihood of the observations under either hypothesis is equal, so Error-statistics inference is indeterminate as to which hypothesis is better. In this case the prior
probabilities directly translate to the posterior probabilities and $H_1$ is chosen as the best hypothesis. Thus Bob’s guilt rests solely on the fact that his explanation for the observations was more convoluted than the simple explanation that he is cheating.

Finally we come to our last example of abduction — that of finding an optimal polynomial curve that fits empirical data. We have limited ourselves to only two possible polynomials (first or second order), and for each of them, two possible choices of coefficients per polynomial. This gives a total of four possible hypotheses to be evaluated given three observed data points. Higher prior probabilities are given to the lower order polynomial due to its shorter syntactic description (elegance) and the fact that it only postulates the existence of two coefficients instead of three (parsimony). The likelihood is highest for $H_4$, the higher order polynomial with coefficients that allow it to fit all three data points at the same time. The other second order polynomial only fits one data point (hence the lowest likelihood), and the two first order polynomials fit two data points, each (hence the same lower likelihood for each of them). Error-statistics inference would then choose $H_4$ as the best hypothesis. Combining likelihoods with expectedness and prior probabilities in Bayes rule leads to the same conclusion in this example. Thus for these particular observations and likelihoods a more complicated hypothesis that more accurately fits more observations is a better choice than a simpler one.
Table 4.1: Posterior and prior probabilities, along with likelihood and expectedness values for the case of Ptolemaic versus Copernican astronomy. Graphically represented in Figure 4.1.

|   | \(P(H|\neg T_1 \land T_2 \land T_3)\) | \(P(H)\) | \(P(\neg T_1 \land T_2 \land T_3|H)\) | \(P(\neg T_1 \land T_2 \land T_3)\) |
|---|--------------------------------|---------|---------------------------------|---------------------------------|
| \(H_1\) | 0.02539 | 0.5 | 0.009801 | 0.1930005 |
| \(H_2\) | 0.97461 | 0.5 | 0.3762 | 0.1930005 |

Figure 4.1: Posterior probabilities on the hypotheses for Ptolemaic versus Copernican astronomy (from Table 4.1). Red stars indicate the theorems that were observed. The deductive tree from Figure 3.1 is ‘pruned’ to remove all implications that do not lead to observations.
Table 4.2: Posterior and prior probabilities, along with likelihood and expectedness values for the case of diagnosing Parkinson-like symptoms. Graphically represented in Figure 4·2.

|   | $P(H | \neg T_1 \land T_2)$ | $P(H)$ | $P(\neg T_1 \land T_2 | H)$ | $P(\neg T_1 \land T_2)$ |
|---|--------------------------|-------|-----------------|------------------|
| $H_1$ | 0.06916 | 0.05 | 0.02 | 0.01446 |
| $H_2$ | 0.58506 | 0.94 | 0.009 | 0.01446 |
| $H_3$ | 0.34578 | 0.01 | 0.5 | 0.01446 |

Figure 4·2: Posterior probabilities on the hypotheses for the diagnosis of Parkinson-like symptoms (from Table 4.2). Red stars indicate the theorems that were observed. The deductive tree from Figure 3·2 is ‘pruned’ to remove all implications that do not lead to observations.
Table 4.3: Posterior and prior probabilities, along with likelihood and expectedness values for an alternate case of diagnosing Parkinson-like symptoms. Graphically represented in Figure 4·3.

|       | $P(H \mid T_1 \wedge \neg T_2)$ | $P(H)$ | $P(T_1 \wedge \neg T_2 | H)$ | $P(T_1 \wedge \neg T_2)$ |
|-------|-------------------------------|--------|----------------------------|-----------------|
| $H_1$ | 0.27894                       | 0.05   | 0.72                       | 0.12906         |
| $H_2$ | 0.72106                       | 0.94   | 0.099                      | 0.12906         |
| $H_3$ | 0                             | 0.01   | 0                          | 0.12906         |

Figure 4·3: Posterior probabilities on the hypotheses for the diagnosis of Parkinson-like symptoms (from Table 4.3). Red stars indicate the theorems that were observed. The deductive tree from Figure 3·2 is ‘pruned’ to remove all implications that do not lead to observations.
Table 4.4: Posterior and prior probabilities, along with likelihood and expectedness values for the case of determining whether or not Bob is cheating on his wife. Graphically represented in Figure 4·4.

Table 4.4

|   | \(P(H|T_1 \land T_2 \land T_3)\) | \(P(H)\) | \(P(T_1 \land T_2 \land T_3|H)\) | \(P(T_1 \land T_2 \land T_3)\) |
|---|----------------------------------|--------|-------------------------------|-----------------------------|
| \(H_1\) | 0.75 | 0.75 | 0.0002 | 0.0002 |
| \(H_2\) | 0.25 | 0.25 | 0.0002 | 0.0002 |

Figure 4·4: Posterior probabilities on the hypotheses that Bob is or isn’t cheating on his wife (from Table 4.4). Red stars indicate the theorems that were observed. The deductive tree from Figure 3·3 is ‘pruned’ to remove all implications that do not lead to observations.
Table 4.5: Posterior and prior probabilities, along with likelihood and expectedness values for the case of finding a polynomial curve to fit empirical data. Graphically represented in Figure 4·5.

|   | $P(H | T_1 \land T_2 \land T_3)$ | $P(H)$ | $P(T_1 \land T_2 \land T_3 | H)$ | $P(T_1 \land T_2 \land T_3)$ |
|---|---|---|---|---|
| $H_1$ | 0.115163 | 0.375 | 0.0004 | 0.0013025 |
| $H_2$ | 0.115163 | 0.375 | 0.0004 | 0.0013025 |
| $H_3$ | 0.001919 | 0.125 | 0.00002 | 0.0013025 |
| $H_4$ | 0.767154 | 0.125 | 0.008 | 0.0013025 |

Figure 4·5: Posterior probabilities on the hypotheses of which order polynomial and which coefficients best predict observed empirical data (from Table 4.5). Red stars indicate the theorems that were observed. The deductive tree from Figure 3·4 is ‘pruned’ to remove all implications that do not lead to observations.
Chapter 5

Conclusions

My first goal in this work was to combine deduction, induction, and abduction into a single coherent framework for inference. I began with deductive inference, which gave us the structure of a formal logical language (FOL in our case), along with its syntax and rule of inference, Modus Ponens. Inductive inference added the concept of probability and thus created a probabilistic logic. Abduction to a best hypothesis was then implicit in this new framework via Bayes rule.

In examining deductive inference, I argued that mathematics and science have the same basic structure, since they both can be represented with FOL, and just differ in their semantics. Making the connection between the two required mapping the concept of ‘observation’ from science to mathematics. Whereas in science, observations are phenomena that we perceive with our senses, I argued that in mathematics the equivalent concept is a numerical relationship that is observed to hold for all empirically tested values. In this view such an observation would eventually be derived from the existing premises, or else be included as a new premise (if independent of the existing ones), or else retain it’s status as a “conjecture”. I realize that this is a very speculative claim, and I have not laid out a convincing air-tight argument for it. Yet it seems that the formal language of FOL provides a means to prove the deductive results of both science and mathematics, which implies some kind of common structure. Several of the most famous logicians and mathematicians of the past century have viewed the increase of mathematical knowledge as a quasi-empirical process, and
thus — I argue — my claim is worthy of further investigation.

When examining inductive inference, I argued that the problem of induction should be recast not just as the problem of creating and justifying hypotheses and implications, but also as the problem of justifying the probabilities assigned to them. I presented various historical theories that provide a rationale for doing this, and made a case for a modified version of Popper’s Propensity theory as the most general and viable method. My small modification drew the idea of assigning probabilities to sentences (rather than to ‘events’ or ‘sets’) from Keynes’ Logical theory.

Upon assigning probability to sentences, mathematics could be seen as the deterministic limiting case of a probabilistic deductive system. This, however, led to a potential problem if probabilities were assigned to the axioms of ZFC. My conclusion was that, since probabilities other than 0 or 1 lead to a paradox, then they are the only values that can be rationally assigned to these particular axioms. This is another speculative idea, and will require much further investigation to determine if the concept of probability is allowable in the context of a (usually) completely deterministic method of inference.

The second goal in this work was to show that the Bayesian method could handle any problem of abduction that contained a finite number of premises and theorems. I attempted to do this by addressing two traditional criticisms of Bayesian confirmation: the Problem of the Catch-all Hypothesis and the Problem of Auxiliary Hypotheses.

I used four examples to illustrate my points throughout this work. They came from the philosophy of science, medicine, law, and statistics. I chose them to be as different from each other as possible in order to show scope of my probabilistic viewpoint. In these examples I assigned probabilities to hypotheses and implications arbitrarily. I argued that the particular values were irrelevant; I only wished to show how probabilities of theorems and posterior probabilities of hypotheses could be
calculated once the values were assigned. Indeed, for cases such as Copernican versus Ptolemaic astronomy, it may be impossible to assign such probabilities. Although not practical in that case, I argued that the Bayesian method can still be applied in principle, and is the correct way analyze the decision-making process of choosing one hypothesis over another.

The exposition of all the topics I presented in this work was brief and cursory; a full examination any one of them would take considerably much more time and effort. Rather than on focusing on just one area, I was more interested in connecting the dots and seeing how a coherent whole that captures all aspects of inference might emerge from probabilistic logic. It is my hope that this work might be a useful clarification of all the concepts discussed.

5.1 Future Directions

There were several topics related to probabilistic logic that were either mentioned briefly in this work, or else not at all, but are very interesting and worthy of further exploration:

1. The one problem of induction that I did not address was how to create hypotheses, or subjectively assign specific values of probability to them. I stated that I consider this outside of the realm of any formal symbolic representation, and relies on some kind of inspiration or access to an objectively real Platonic world of ideas. It therefore lies in the realm of psychology, neuroscience, or theology, and was thus outside of the scope here. Any further study of this would then require looking into one, or all, of those areas.

2. Related to the first topic, the idea of syntactic elegance as an objective method for assigning probabilities is an extremely important topic. If feasible, it would completely transform practical inference methods. There would no longer be
any subjectivity in the process of assigning probabilities other than using human intuition to try and find the minimum length representation of any given formula (and thus its probability), since this is known to be an uncomputable task.

3. The paradox from Section 3.2.8 was intriguing, and it would be interesting to see if other deductive systems similarly limit the values of probability that can be assigned to their axioms. A place to start would be to examine the logical axioms themselves, such as, e.g., the Law of Excluded Middle.

4. Robustness has recently become a hot topic in philosophy (Levins, 1966; Orzack and Sober, 1993; Weisberg, 2006; Calcott, 2011; Kuorikoski et al., 2010). The word can refer to a number of different concepts, but what they have in common is the idea of stability under some kind of perturbation. In a non-probabilistic deductive setting a robust theorem is one that is implied by multiple different hypotheses and is stable under perturbation of those hypotheses — where ‘perturbation’ in this case refers to the removal of one or more of them. In a probabilistic logic robustness would then involve not just the number of hypotheses and implications, but also the probabilities assigned to them. It would be interesting to see whether robustness offers an alternative to the Bayesian method for determining the most probable theorems and abduced hypotheses.

5. I did not examine how inference within probabilistic logic changes when there are countably, or uncountably infinite numbers of sentences involved. It was beyond the scope here, and a full analysis of the would require an in-depth study of the justifiability and feasibility of the concept.

6. Whereas the aspects of simplicity known as elegance and quantitative parsimony could be shown to fit in my probabilistic logic framework, I did not examine how qualitative parsimony (i.e., the number of different kinds of entities being
postulated) enters into it. It is currently not clear how more types of entities would lower the probability of a hypothesis.
References


