University of Amsterdam

MSc Physics
Theoretical Physics

Master Thesis

On the Cox-Jaynes justification for objective Bayesian probability theory and the mind projection fallacy in physics

---

by

Tim Bakker
10177302

June 2016
60 ECTS
14-09-2015 to 14-06-2016

Supervisor:
dr. Peter Grünwald

Examiners:
dr. Ben Freivogel
dr. Christoph Weniger

Institute of Theoretical Physics
Faculty of Natural Sciences, Mathematics and Computer Science

Abstract

We discuss Cox's theorem as a justification of objective Bayesian probability theory. We note the primary points of controversy in the literature - the representation and density axioms - and attempt to elucidate the current state of the theorem. Recent work by Alexander Terenin and David Draper is examined that allows for a resolution to the controversy surrounding the density axiom, and that provides a connection between Bayesian methods and Kolmogorov's influential probability axioms. Edwin Jaynes's extension of Cox's theorem to 'probability theory as logic' is moreover discussed, as is the related concept of the mind projection fallacy: a cognitive bias that amounts to confusing ontic and epistemic properties of nature. The consequences of this fallacy for interpretations of probabilities are considered, with particular focus on the physical theories of classical statistical mechanics and quantum mechanics.

Contents

Chapter 1

Introduction

In 1946 Richard Cox wrote his seminal paper discussing systems of plausible reasoning (Cox 1946). He proposed a number of intuitively-appealing axioms and consequently showed that those systems consistent with his axioms are isomorphic to probability theory. This statement - known as Cox's theorem - was the first attempt to formalise the Bayesian approach to probability theory based on reasonable expectation (Terenin and Draper 2015). Previous attempts to formalise probability theory had been made by Bruno de Finetti (1931, 1937) and Andrey Kolmogorov (1933). The former based his approach on subjective degrees of belief, while the latter was concerned with finding a minimal set of axioms that enabled the proving of all (then) known standard theorems of probability theory. Cox's theorem was extended by Edwin Jaynes (1968, 1990, 2003) and in its current form the theorem is known as the Cox-Jaynes (CJ) approach to Bayesian probability theory.

The CJ approach is interesting for a number of reasons. First of all, it provides a mathematical justification for Bayesian probability theory as an objective theory of one-dimensional plausibility: it shows that Bayesian probability is the proper method for making decisions under uncertainty. In subjective Bayesian probability this was already known as the coherence of bets in the context of de Finetti's work (1931, 1937). Cox's theorem however provides this justification for the objective case, where one's plausibility assignments are uniquely determined by the available information. An introduction to Bayesian probability theory, as well as a discussion of its subjective and objective perspectives is provided in Chapter 2.

Secondly, Cox's theorem allows a connection to be drawn between Bayesian probability theory and the probability set-axioms of Kolmogorov. Kolmogorov's axioms have long been considered the standard axioms of probability theory, and attempts to reconcile them with Bayesian axiomatisations have historically encountered the issue that the latter does not support a countably additive probability function. While Cox and Jaynes themselves have not addressed these problems sufficiently, a recent work by Alexander Terenin and David Draper (2015) provides a possible solution. The reconciliation of Cox's theorem with the Kolmogorov axioms represents a strengthening of the objective Bayesian perspective's foundations. Furthermore, the work by Terenin and Draper addresses another controversy surrounding the CJ approach; that of the proposition space's cardinality. Chapter 3 is dedicated to the derivation of Cox's theorem, while Chapter 4 discusses the the aforementioned controversies and Terenin and Draper's contribution.

A final point of controversy in the CJ approach - also discussed in Chapter 4 - is the representation axiom, which states that one's reasonable belief should be represented by

a single real number. Objective Bayesian methods following this axiom have historically been criticised for being unable to adequately manage situations of extreme ignorance, where very little information is available. As an example, a probability assignment of 0.5 to a flipped coin coming up heads may indicate both the belief that the coin is definitely fair, or the complete ignorance belief that there is no information available about the coin, beyond the fact that it is two-sided. The inability to distinguish between these two situations is reason to doubt the applicability of objective Bayesian methods.

Appendix B examines some further arguments against the objective Bayesian view based on this failing. Theories of probability that represent information by two real numbers - i.e. two-dimensional theories, such as imprecise probability theory, discussed in Appendix B.2 - generally do not share this problem, distinguishing between the two situations using their second degree of freedom. As we will see, the criticism of the representation axiom remains valid throughout our discussion, and as such the current limitations of the objective Bayesian formalism should be recognised.

The final reason for focusing on the CJ approach is that it provides a stepping stone for the introduction of Jaynes's theory of 'probability as extended logic' (1989; 1990; 1996; 2003). Cox's theorem proves that objective Bayesian probability is the unique proper system for reasoning under uncertainty, assuming one accepts his basic axioms. Jaynes viewed this proof as an opportunity to extend Bayesian probability theory to apply to every case of logical inference, so that one may make consistent logical statements even in situations where information is missing. In this way, he viewed Bayesian probability as an extension of classical deductive logic to cases where deduction is impossible due to unavailability of relevant information.

Jaynes's view of probability theory as logic has resulted in a number of remarkable achievements. He developed the methods of entropy maximisation (Appendix A.2) and transformation groups (Appendix A.3) for constructing objective prior distributions, necessary for objective Bayesian inference. Perhaps most notably, these methods allowed him to solve the Bertrand paradox, a problem that was often considered unsolvable due to being underspecified (Jaynes 1973).

The success of probability theory as logic lead Jaynes to the extreme objective Bayesian view that all probabilities must represent some measure of reasonable belief, and are thus never indicative of actual stochastic processes in nature (Jaynes 1990). According to him, the belief that probabilities are actually physical is a form of the mind projection fallacy, which more generally refers to a confusion of ontic (relating to being) and epistemic (relating to knowledge) properties of nature. This claim and its consequences for interpretations of probabilities are discussed in detail in Chapter 5.

Particular focus will rest on interpreting probabilities in the physical theories of classical statistical mechanics and quantum mechanics. Jaynes argues that the former is essentially a theory of inference from incomplete information, rather than a theory describing statistical behaviour in physics (Jaynes 1957a; Jaynes 1957b; Jaynes 1989; Jaynes 1990). Indeed, while Jaynes's arguments for the classical case seem unassailable (Nathaniel 2012), the extension to quantum mechanics faces a number of problems.

In particular, Jaynes argues that Bell's original no-go theorem (Bell 1964) and the subsequently performed experiments (Freedman and Clauser 1972; Aspect et al. 1981; Aspect et al. 1982a; Aspect et al. 1982b; Abellán et al. 2015) do not completely exclude local deterministic hidden variable theories, as is commonly believed (Jaynes 1989). Our current quantum mechanical formalism may then be a theory of inference from incomplete information of these hidden variables. Roger Colbeck and Renato Renner (2011) agree

with much of Jaynes's criticism of Bell, but propose a strengthening of Bell's theorem that shows no hidden variable theory (either local or non-local) can produce more informative predictions than current quantum mechanics; thus excluding most of Jaynes's arguments that do not invoke superdeterminism.

There may remain one possible method to salvage Jaynes's position of quantum mechanics as inference, stemming from difficulties in interpreting probabilities in current quantum mechanics as epistemic. Jaynes notes that such an interpretation is not as simple as in classical statistical mechanics, since quantum probabilities do not contain all the information available in the wavefunction from which they are derived (Jaynes 1996b). He then considers the possibility that current quantum probabilities are not what a theory of inference should reproduce, instead postulating a deeper hypothesis space that does represent all of the experimenter's information, which should allow for more informative inferences than the current formalism does.

As we shall see, the case for quantum mechanics as a theory of inference is not nearly as strong as that for statistical mechanics. However, there may be sufficient reason to further explore the possibility, as such research may advance our understanding of quantum mechanical occurrences, such as wavefunction collapse.

The purpose of this thesis is then to provide a modern overview of Jaynes's views. A number of the aforementioned works - such as those by Terenin and Draper, or Colbeck and Renner - were published after Jaynes's death in 1998, but have interesting implications for his ideas on probabilities. We have endeavoured to combine these different works with his existing views into a coherent whole, both strengthening and weakening Jaynes's various positions with the insights of these later authors. It is our hope that this thesis encourages interest in ideas of probability as a measure of information, and particularly in the consequences of such ideas for physical theories.

Chapter 2

An introduction to Bayesian probability theory

It seems appropriate to first write an introductory note on the Bayesian approach to probability theory itself. Since teaching on probability theory is still mostly focused on the frequentist approach (Bernardo 2006), the first section of this chapter is devoted to contrasting the frequestist and Bayesian perspectives. We will provide a short introduction on frequentist methods and philosophy in section 2.1.1 and a slightly more extensive overview of Bayesian perspective in section 2.1.2. The latter moreover serves to introduce Bayes's theorem and the use of prior probabilities. Section 2.1.3 briefly explores the differences between frequentist and Bayesian perspectives in a more nuanced manner. This will be followed in section 2.2 by a short overview of the history of Bayesian probability, to illustrate its historical significance and current popularity in academia. We will end the chapter with a discussion of the two philosophical schools within the Bayesian interpretation; those of objective and subjective Bayesianism in section 2.3.

2.1 Frequentist and Bayesian probability

Consider flipping a fair coin. The probability of such a coin landing on heads is well known to be a half. This is true because if the experiment of flipping the coin were to be repeated an infinite number of times, the ratio of the number of times a specific outcome occurred and the number of times the coin was flipped, would be $\frac{1}{2}$. The probability of an event x that occurs n_x times out of n_t total experiments is then given by

$$P(x) = \lim_{n_t \to \infty} \frac{n_x}{n_t}. \tag{2.1}$$

This interpretation of probability - called the frequentist interpretation - had until recently been the dominant one for over a hundred years: after its introduction by mainly Ellis and Cournot (Ellis 1843; Ellis 1854; Cournot 1843) and criticism on the then widespread classical definition of probability by Venn and Boole (Gigerenzer et al. 1989), the frequentist interpretation emerged as the main body of statistical tools for empirical sciences, due to its application to repeatable objective processes inherent in scientific experiments.

2.1.1 Frequentist methods and philosophy

In order to discuss frequentist methods, it is important to note the main consequence of interpreting probabilities as expressions of limiting frequencies over a large number of experiments; namely that a frequentist view cannot easily state conclusions about single events. Frequentist probability makes statements about a large population of equivalent events: that is, statements about sampling from populations. Indeed, this is the only way to obtain limiting frequencies in the manner of equation (2.1). Applying this to our example of flipping a coin, we may consider probabilistic statements about a random variable that represents the coin in some sampling experiment, but not directly about a single coin flip.

In the context of model selection frequentists are often concerned with maximising the likelihood $L(\theta|X)$ of a model parameter θ given a sample of data X. This (for simplicity discrete) likelihood is defined as $L(\theta|X) = p(X|\theta)$; the likelihood of θ given X is the probability of X given θ. Maximising the likelihood is akin to finding the parameter θ that is most likely given the data X (Vaughan 2013). From this it is clear that the parameter θ is considered a fixed property of the system: it is the data that is random.

This particular viewpoint is also manifest in the interpretation of the confidence intervals often used in parameter estimation: a 95% confidence interval on some parameter θ is an interval that will contain the true value of the parameter 95% of times if the exact same experiment were to be repeated many more times (to be precise: an infinite number of times) (Cox and Hinkley 1974). It is the bounds of the interval - the data - that are the random variables here; the parameter itself is fixed by the limiting frequencies of the process being studied. The interval is thus not one that has a 95% chance of containing the true parameter value in any one instance of the experiment; as before, the frequentist interpretation can make no such statements (Kendall and Stuart 1973).

2.1.2 Bayesian methods and philosophy

In such situations the Bayesian interpretation of probability theory, introduced in this section, is more useful. It views probabilities as statements not about the nature of a system, but about one's belief in propositions about and ignorance of that system. For Bayesians the $\frac{1}{2}$ probability associated with a coin is not an attribute of the coin necessarily. Rather, it shows one's belief that the coin has probability half to come up on either side; i.e. the coin is fair.

This is however merely a description of one's belief state, since were one to know the exact force with which the coin was thrown, the exact movements of molecules in the air it travels through, the exact configuration of forces on the table in lands on, and a multitude of other relevant properties, one could in principle determine the exact face the coin would land on and not require any probabilities to describe the system. The additional knowledge has thus changed the relevant probabilities. Indeed, it is in principle possible to program a robot that flips coins in such a way that they always land on heads. It is only the lack of knowledge of relevant forces that leads one to suppose the flip is random, and should be described by probabilities.

This interpretation allows a Bayesian to make statements about single events, such as the probability that a specific coin lands on heads. As such the Bayesian interpretation may be viewed as more broadly applicable than the frequentist one. Bayesian methods however, are often much more difficult to effectively implement that frequentist ones, which may explain the latter's historical prominence (see also section 2.2).

The Bayesian equivalent of confidence intervals are credible intervals, which have exactly the interpretation that the frequentist confidence intervals were mentioned to have not. A 95% credible interval is an interval that contains the true parameter value in 95% of all cases. Here it is the parameter that is the random variable - as opposed to the interval's bounds (the data) - which is an expression of the Bayesian experimenter's ignorance about that parameter.

Bayesian probability is concerned with updating one's belief in propositions or the accuracy of one's models by gathering data. A Bayesian has to first posit some distribution that represents either their belief in a proposition, or arguments based on reasonable plausibility (see section 2.3); this is called the prior distribution, or simply prior. For instance, a prior on our coin may be that the coin is fair, which then assigns $\frac{1}{2}$ probability to each side. Such priors may be either subjective or objective, depending on the Bayesian's school of thought (section 2.3 explores these differing views in more detail). The possible subjectivity is often cause for criticism of the Bayesian interpretation (Jaynes 1990).

After determining a prior one should gather data; i.e. flip the coins. After tallying the number of times a specific side came up one might for instance notice a significantly higher number of heads. The prior distribution can then be updated using this data to form the posterior distribution describing one's belief about the coin. This posterior distribution will assign a higher probability to heads coming up, in a way that is consistent with the observed data. Updating happens via Bayes's theorem, from which the Bayesian interpretation obtains its name. In the simplest form Bayes's theorem is (Fienberg 1992):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{2.2}$$

Here $P(A)$ represents the prior information, in this case that the coin is fair. $P(B|A)$ is the term representing the data: it may be interpreted as the probability of some experimental outcome (B) occurs given that the coin is fair (A). This term is often identified as the likelihood from frequentist statistics: here the model parameter θ could represent for instance the hypothesis A that the coin is fair, or the probability of the coin landing on heads. As opposed to the frequentist, the Bayesian considers θ variable as it represents their ignorance about the coin.

$P(A|B)$ is the posterior distribution and represents the updated belief: the probability that the coin is fair given the observed data. Finally, $P(B)$ essentially serves as a normalisation factor, but may be interpreted as the probability that the data occurs in isolation of any other knowledge.

Bayes's theorem can be applied to either distributions - representing models - or point probabilities - representing a measure of belief in a given proposition. In the example above both cases have been used interchangeably to illustrate this: A can either represent the distribution of a fair coin, one's belief that the coin is fair, or the probability one assigns to the coin coming up heads. Similarly, $P(B|A)$ may be the distribution of the data B given a prior distribution A, or the probability that B is observed given that the coin is fair, or comes up heads with a given prior probability. Whichever application is chosen depends on one's needs; the posterior distribution will also either end up being a distribution or a point probability.

2.1.3 Reconciliation of perspectives

It may now seem that frequentists and Bayesians have strongly differing opinions of what a probability represents. In this section we will attempt to briefly show that there need not necessarily be conflict; indeed, there may be multiple avenues for reconciliation between the two perspectives.

Viewing probability as a statement of knowledge is integral to the Bayesian philosophy, but this is not to say that the frequentist philosophy is necessarily wrong from the Bayesian perspective. Consider again the example of flipping a coin. When frequentists state the probability of a fair coin coming up heads is a half, they are simply saying that on average - perhaps after an infinite number of coin flips - the frequency of heads to total flips will be a half. This is certainly true for a fair coin, unless thrown by a biased system such as the robot programmed to always flip heads. In that case, the frequentist probability of the coin coming up heads will be one, and indeed the frequency of heads to total flips will also be one.

The only difference here is in interpretation: Bayesians do not necessarily associate frequencies with probabilities, since the Bayesian viewpoint wants to be able to make statements about single events. However, the Bayesian probability (i.e. belief or knowledge) should in the long run correspond with the frequentist probability (frequency), as those frequencies are measurable quantities. To Bayesians a probability is a theoretical construct that represents a state of knowledge, while a frequency is a measurable property of the real world. Probability does not tell a Bayesian how the world is, but rather it is a mathematical tool for ensuring the consistency of one's own reasoning (Jaynes 1989).

In essence, one may choose to say that probabilities are frequencies, but that seems needlessly narrow to Bayesians. Instead, probability describes information, while frequency describes experimental outcomes. With sufficient information (i.e. the results of experiments) the two will be equivalent, but probabilities have wider applications as measures of prior beliefs. The Bayesian perspective on probability may be best understood as the aphorism: 'every frequency is a probability, but not every probability is a frequency'.

There exists another possible conflict, however. From the frequentist perspective, the possible subjectivity of a chosen prior is often cause for concern. Indeed, for example in the problem of model selection, one could imagine that choosing a prior far from the actual distribution could lead posterior distributions almost equally far from the true distribution. To assuage this concern, we note the fact that it has been shown that choosing a prior far from the actual distribution is not always problematic in model selection. As long the model being evaluated satisfies certain properties - such as smoothness and finite-dimensionality - and as long as enough data is gathered, the Bernstein-von Mises theorem guarantees that the frequentist and Bayesian methods yield approximately the same results (Freedman 1999). For instance, the Bayesian posterior is close to the frequentist maximum likelihood estimate for a large enough sample. Since frequentist statistics uses no prior distributions the theorem implies that posteriors are almost independent of priors when updated with enough new data (van der Vaart 2000). This result provides a basis for reconciliation between the two approaches (Boucheron and Gassiat 2009), and is quite important in showing the validity of the Bayesian approach, since Bayesian probability should in the long run correspond with the measurable frequencies.

While both frequentist and Bayesian approaches possess their share of technical issues - for instance the dependence of a p-value on unobserved data and subjective intentions (van der Pas 2010) for the former and subjectivity of priors for the latter - they nonetheless have widespread applications and their results may often overlap. Frequentist methods are often

easier to apply to idealised versions of real world problems, while Bayesian methods have in the last few decades gained popularity due to their flexibility, ubiquitous applications and increased computing power (Malakoff 1999; Jaynes 2003).

A for physicists especially interesting argument in favour the Bayesian philosophy has been provided by Edwin Jaynes (1990; 1989). Jaynes was a proponent of what he called 'probability theory as logic', in which Bayesian probability theory is extended to apply to every case of logical inference, so that one may make consistent logical statements in situations where information is missing. This in a sense extends classical deductive logic to probabilistic cases (Jaynes 2003). Jaynes claims probabilistic arguments in physics - and *"every field where probability theory is used"* (Jaynes 1990, p.2) - have long been plagued by the mind projection fallacy; a lapse in reasoning where one asserts that one's own state of mind is indicative of real properties of nature. Specifically, he refers to the common belief that probability distributions used in classical statistical mechanics and quantum physics represent actual stochastic - as opposed to deterministic - behaviour in nature. Chapter 5 is dedicated to an exploration of these controversial claims.

2.2 A brief history of Bayesian probability

In section 2.1.2 it was mentioned that Bayesian probability obtains its name from Bayes's theorem. The theorem in turn is named after Thomas Bayes, who proved a special version in the paper *An Essay towards solving a Problem in the Doctrine of Chances*, posthumously published in the year 1763 (McGrayne 2012). Early Bayesian inference was further developed by Laplace who derived the general theorem and applied it to a variety of problems at the time, including celestial mechanics, medical statistics and law (Stigler 1986). These methods, sometimes called 'the calculus of inductive reasoning' are the forebearers of what Jaynes calls 'probability theory as logic', and indeed Laplace used them as a generalisation to Aristotle's deductive reasoning in situations of incomplete information (Jaynes 1996b). Laplace utilised what he called the 'principle of insufficient reason' (also called 'principle of indifference') to determine the priors required to solve his problems of Bayesian inference. It is equivalent to assigning equal probabilities over a set of propositions, and is the simplest objective prior given no other information.

These methods were summarised under the moniker of 'inverse probability', as they infer backwards from observed data to the random parameters. Considered not general enough, they were replaced by frequentist statistics at the start of the twentieth century (Fienberg 1992). Particularly the principle of indifference was deemed too narrow, as it could only be invoked in the presence of natural symmetry - such as with coins or dice - and such symmetry was not always present in problems of the time (Gigerenzer et al. 1989).

The theory thus required further expansion to be relevant, and such development came in the form of Jeffreys's 1939 book *Theory of Probability* (1939). Jeffreys proposed a number of priors that worked for problems where indifference was not tenable, and is often credited with the revival of the Bayesian viewpoint (Appendix A.1). Contemporary works by de Finetti (1931; 1937), Cox (1946), Wald (1949) and Savage (1954) either expanded on Jeffreys's achievements, or discovered ways of supporting the - either objective or subjective - Bayesian perspective with arguments from plausibility and decision making.

Despite this, it was not until the 1980s that Bayesian methods fully returned to academia. Frequentist methods were generally considered more objective, and were even preferred to objective Bayesian approaches due to the latter's myriad of computational difficulties. The increasing prevalence of ever more powerful computers and discovery of improved

computational techniques - such as Markov Chain Monte Carlo methods - lead the way for Bayes's return to relevance (Wolpert 2004). Today Bayesian methods are widely accepted and used in fields such as physics, machine learning, medicine, law, artificial intelligence and many more.

2.3 Objective and subjective priors

In the previous there have been several mentions of two types of priors in Bayesian statistics; those that are subjective and those that are objective. Adherents of using a specific type of prior over the other are similarly called respectively subjective and objective Bayesians. These latter terms only emerged around the year 1950, but have since become standard nomenclature (Fienberg 2006).

The differences between these types of Bayesian lie partially in the interpretation of probabilities, and partially in methodology. Probabilities constitute measures of belief from both Bayesian viewpoints, be it the belief that a given hypothesis is true, or the belief that a coin will come up heads. To a subjective Bayesian these probabilities are personal; they represent an person's belief that a certain proposition is true, or that a certain event will happen, given the knowledge available to that individual. Two subjective Bayesians can come to different conclusions regarding probabilities, even if they appear to have the same relevant information; one may for example evaluate a certain piece of information differently than the other, for whatever reason, which leads to different assignments of belief. When designing an experiment, a subjective Bayesian may also ask an expert in the field for a prior based on the expert's knowledge.

An objective Bayesian however views probabilities as representing a degree of reasonable belief. This degree should be uniquely determined by the available information, and not contain any subjective components; two objective Bayesians should in principle come to the same conclusion given the same information. Objective Bayesians have rules for determining priors that represent these reasonable belief states. These rules include the Jeffreys's rules (Jeffreys 1931; Jeffreys 1939; Jeffreys 1955; Jeffreys 1946 and Appendix A.1), maximum entropy methods (Jaynes 1957a; Jaynes 1957b; Jaynes 1963 and Appendix A.2), prior transformation groups (Jaynes 1968 and Appendix A.3), reference analysis (Bernardo 1979; Berger and Bernardo 1991) and a multitude of others (Kass and Wasserman 1996; Ghosh 2011).

Both approaches have their own advantages and disadvantages. The subjective approach is easier to apply, but selecting a prior too far from the true distribution may lead to posteriors that do not accurately capture this distribution. In general proper selection of priors becomes less important the more data is collected; nevertheless, proper prior selection remains important because data is limited in practice. On the other hand, the objective approach necessarily uses a prior that encompasses all the information an experimenter has (and no more). Devising such a prior however is not easy except in the simplest of cases; for instance, those where the indifference prior can be used due to symmetry considerations.

As mentioned in the previous section, attempts have been made to mathematically justify both Bayesian perspectives. Bruno de Finetti (1931; 1937) famously made such an attempt for the subjective approach in the context of betting. He showed that one cannot be exposed to guaranteed losses by an array of compulsory bets when one's assignment of odds in those bets follows the rules of Bayesian probability. It is then said that such a bettor is immune to being 'Dutch booked'. An assignment of bets that follows this scheme

is called coherent; coherence is the central principle of de Finetti's work for justifying Bayesian probability.

Leonard Savage (1954) showed somewhat similarly that a rational decision maker should act in accordance with Bayesian rules if they are to maximise their subjective utility. Utility is a term that originates in economics and decision theory and may be interpreted a measure of a decision maker's satisfaction with a certain outcome. Maximising this utility involves making decision in the presence of incomplete information, and consistency requires handling such uncertainty through the use of Bayesian probability theory.

Abraham Wald (1949) proved that every unique Bayesian procedure is admissible, and vice-versa, that every admissible procedure is Bayesian. Admissibility is a property of decision rules; an admissible decision rule is one such that no other rule is always better than it, where 'better' refers to one's utility. Admissibility is the decision theoretic analog of Pareto optimality; a state in which no strict improvements to a system can be made by a different allocation of resources. Thus, Wald showed there is no strictly better way of making decisions than through Bayesian probability theory.

A thorough examination of these works is surely interesting, but beyond the scope of this thesis. Instead our focus lies on a justification for the objective Bayesian viewpoint by Cox, based on reasonable expectation, called Cox's theorem (Cox 1946). Cox proposed a number of somewhat intuitively-appealing axioms and consequently showed that those systems consistent with his axioms are isomorphic to probability theory. This approach, though not without controversy, is out of all the previous perhaps the most interesting for a researcher as it does not rely on subjective representations of belief. For this reason, it is the focus of our next chapter.

Chapter 3

The Cox-Jaynes axiomatisation

This chapter is dedicated to proving Cox's theorem, which may be considered a derivation of a unique coherent system of plausible reasoning, or reasoning under uncertainty. Starting from - what he considered - natural axioms, Cox was able to show that such a system must be isomorphic to probability theory. Cox's original argument has been heavily supported and extended by Jaynes (1968, 1990, 2003), which has earned it the name 'Cox-Jaynes (CJ) approach'.

The CJ approach starts from the viewpoint that any proposition can be assigned some value representing reasonable belief, making it a justification of Bayesian probability. Cox's theorem was the first attempt to formalise the Bayesian approach based on reasonable expectation (Terenin and Draper 2015). In particular, the CJ approach is often regarded as a support for the objective school of Bayesianism, in which some objective plausibility must be uniquely determined for every proposition by the available information.

In this chapter we shall adopt the notation and definitions employed by Kevin Van Horn (2003). We follow his derivation of Cox's theorem closely in the following. While Van Horn's proof is not quite equivalent to either that of Cox or Jaynes, it paints a clearer picture of the exact - controversial - state of the theorem. Jaynes provides a more extensive 'common sense' proof in Chapters 1 and 2 of his 2003 book *Probability Theory: The Logic of Science*, parts of which will appear below as well.

In the following, we will first set forth some definitions in section 3.1, after which the basic axioms of Cox's theorem will be introduced in section 3.2. Next, in section 3.3, we will derive the rules of probability theory from these axioms, concluding the proof of Cox's theorem.

As one may well find during a reading of section 3.2, not all of the presented axioms are equally evident. Cox's theorem has garnered some controversy for this reason, which will be explored in Chapter 4.

3.1 Basic definitions

As the title suggests, this section is dedicated to establishing some definitions. Many of these will be familiar to those with some knowledge of Boolean logic, though some definitions refer to concepts beyond this domain.

Definition 1: *A proposition* is a statement that is either true or false.

Definition 2: *A compound proposition* is a proposition that is constructed from multiple other propositions, using the operators $\neg$ (negation), $\wedge$ (and), $\vee$ (or), $\Rightarrow$ (implies), and $\Leftrightarrow$ (equivalence). An example of a compound proposition (that is trivially true) is $(A \vee \neg A)$, where A is any proposition.

Definition 3: *An atomic proposition* is a proposition that is not compound. For instance, 'y = a' is an atomic proposition, as is 'grass is green'.

Definition 4: *A state of information* X characterises the information we have about some set of atomic propositions $\mathscr{A}$ and their relation to each other. Here $\mathscr{A}$ is called the *basis* of X. The domain of X is the union of $\mathscr{A}$ and all compound propositions that involve only atomic propositions in $\mathscr{A}$. We call this the logical closure of $\mathscr{A}$.

A state of information contains all information that influences one's assignment of plausibilities to their propositions. It is not only concerned with information of certainty ('$2 > 1$', 'this is a thesis'), but rather also with any statement of uncertain information ('$X \sim U(1, 10) > 5$', 'there's a 10% chance it will rain today').

Definition 5: $A|X$ or $(A|X)$ is the plausibility assigned to A given the state of information X. The state of information obtained by adding to X that proposition A is true is denoted by A, X.

Note that we are not introducing probability theory here, despite the fact that notation for the plausibility $(A|X)$ has striking similarities to the conditional probability $P(A|B)$. No statements have yet been made about how $(A|X)$ should be represented, or how it can be manipulated.

Definition 6: *A tautology* is a proposition that is true regardless of whether its atomic propositions are true or false. For instance, $(A \vee \neg A)$ is a tautology, since it is true whether A is true or false.

Definition 7: Two propositions A and B are equivalent if $A \Leftrightarrow B$ is a tautology. The trivial example is $A \Leftrightarrow A$. Another example is $(2 < 1) \Leftrightarrow (3 < 2)$, since these are both always false.

Definition 8: A state of information X is *consistent* if there is no proposition A for which both $(A|X) = \mathcal{T}$ and $(\neg A|X) = \mathcal{T}$. An example of an inconsistent state of information is thus $(A, \neg A, X)$.

3.2 The axioms

Our definitions established, we now turn to Cox's axioms. The set of axioms given in this section is not the unique one that enables the proving of Cox's theorem; a number of slightly different sets exist (Halpern 1996; Halpern 1999; Terenin and Draper 2015). In particular the axioms as presented here are not identical to those provided in Cox's original paper (Cox 1946). Instead, we follow the prescription of Van Horn (2003), who in turn based his approach on Paris (1994).

Axiom 1: representation. The plausibility of A given X, $(A|X)$ is a real number. There furthermore exists a number $\mathcal{T}$ such that $(A|X) \leq \mathcal{T}$ for all A and all X.

Since most things of any sort of magnitude are measured using real numbers (time, distance, temperature, etc.) the first part of the axiom may seem reasonable. We can interpret the second part to mean that higher plausibilities are denoted by larger numbers, and $\mathcal{T}$ is the value assigned to a proposition that is known to be true. One might think

we would encounter a problem if we were to denote known truth or falsity by $+\infty$ and $-\infty$, since those are not real numbers. However, it is always possible to bijectively map such a representation to a finite interval using a continuous, strictly increasing, invertible transformation, such as $f(x) = \arctan(x)$.

While this axiom may seem reasonable, it is perhaps the most controversial out of all of them. The controversy will be elaborated on in section 4.1 of the next chapter.

Axiom 2: consistency with propositional calculus. Plausibility assignments are consistent with the propositional calculus. This comprises four statements:

1. If A is equivalent to A', then $(A|X) = (A'|X)$.
2. If A is a tautology, then $(A|X) = \mathcal{T}$.
3. $(A|B, C, X) = (A|(B \wedge C), X)$.
4. If X is consistent and $(\neg A|X) < \mathcal{T}$, then A, X is also consistent.

Informally, these may be restated as:

1. If A and A' are equivalent, we should be able to use them interchangeably.
2. Any tautology is known to be true, and our plausibility assignment should reflect this.
3. If B and C are true, then $B \wedge C$ is true. Vice-versa, if $B \wedge C$ is true, then B and C are true.
4. If we cannot say with certainty that knowledge of X renders $\neg A$ true, then A could still be true. Thus A cannot contradict X.

Axiom 3: negation. There exists a nonincreasing function S_0 such that $(\neg A|X) = S_0(A|X)$ for all A and consistent X.

In a very real sense, A and $\neg A$ answer the same question: knowing the truth value of A allows one to determine the truth value of $\neg A$ from the propositional calculus. This axiom is an extension of that to plausibilities; i.e. it also deals with propositions that are not necessarily true or false.

Furthermore, the caveat that S_0 be nonincreasing ensures that a change in state of information does not simultaneously allow A to become more and less plausible. That is, if a state of information changes from X to Y, and $(A|X) < (A|Y)$ then $S_0(A|X) > S_0(A|Y)$, or $(\neg A|X) > (\neg A|Y)$. Hence, if knowing Y renders A more plausible than knowing X, it also renders $\neg A$ less plausible.

Cox himself only explicitly required his equivalent of S_0 be twice differentiable, but careful reading reveals his implicitly assumption that it be strictly increasing or decreasing (Horn 2003, footnote 6). Of course, one could take S_0 to be nondecreasing instead of nonincreasing as well, though this would flip the interpretation of our plausibilities. This would consequently require redefining $\mathcal{T}$ in Axioms 1 and 2 as well.

Finally, note that if $(A|X)$ and $(\neg A|X)$ are allowed to vary independently of each other, then two real numbers would be required to summarise our uncertainty about A. This contradicts Axiom 1, hence any controversy surrounding that axiom also applies to Axiom 3. Let us now consider two immediate consequences of these first three axioms:

Proposition 1: $\mathcal{F}$ and the plausibility domain. Define $\mathcal{F} = S_0(\mathcal{T})$. Then $\mathcal{F} \leq (A|X) \leq \mathcal{T}$ for all A and consistent X.

Proof: $\mathcal{T} \geq (A|X) = S_0(\neg A|X) \geq S_0(\mathcal{T}) = \mathcal{F}$. □

Proposition 2: involution. If X is consistent and $x = (A|X)$, then $x = S_0(S_0(x))$.

Proof: $x = (A|X) = (\neg\neg A|X) = S_0(\neg A|X) = S_0(S_0(A|X))$. □

The next axiom is another one that has garnered much controversy.

Axiom 4: density. There exists a nonempty set of real number $\mathcal{P}_0$ with the properties that:

1. $\mathcal{P}_0$ is a dense subset of $(\mathcal{F}, \mathcal{T})$. That is, for every pair of real numbers a, b such that $\mathcal{F} \leq a < b \leq \mathcal{T}$, there exists some $c \in \mathcal{P}_0$ such that $a < c < b$.

2. For every $y_1, y_2, y_3 \in \mathcal{P}_0$ there exists some consistent X with a basis of at least three atomic propositions A_1, A_2, A_3, such that $(A_1|X) = y_1$, $(A_2|A_1, X) = y_2$ and $(A_3|A_2, A_1, X) = y_3$.

The necessity of this density axiom will become clear in sections 3.3.1 and 3.3.2, but the question remains why one should find this a reasonable assumption. Since the purpose of Cox's approach is to extend propositional calculus to a logic of plausible reasoning, one should expect his theory to at least be able to reason correctly about propositions the propositional calculus can reason about. Furthermore, since propositional calculus is able to reason about any domain for which useful propositions can be devised, the same should be true of this new logic of plausible reasoning. One might then think the logic of plausible reasoning should (at least) be able to handle cases of three unrelated atomic propositions with arbitrary plausibilities.

Note that Axiom 4 requires less than this. Firstly, the requirement is not that the logic should handle arbitrary plausibilities. Rather, it is allowed that some of the real numbers in the interval $[\mathcal{F}, \mathcal{T}]$ are not admitted as plausibilities, but we do require $\mathcal{P}_0$ to be dense, so that the theory has no holes. For example, restricting $\mathcal{P}_0$ to only contain rational or algebraic numbers would still be in accord with Axiom 4.

Secondly, Axiom 4 does not necessitate that the three atomic propositions are unrelated. If that were so, the requirement would be that knowledge about the truth state of a proposition does not change the plausibility of another, thus $(A_i|A_j, X) = (A_i|X)$, $(A_i|A_j, A_k, X) = (A_i|X)$, etc. Axiom 4 does not require this condition; instead, it requires something weaker that is still consistent with that condition. This is readily seen by choosing arbitrary values for several A_i and then applying the equations above.

A consequence of this axiom is that Cox's space of plausibility values is nontrivial.

Proposition 3: non-triviality of the plausibility space. $\mathcal{F} < \mathcal{T}$.

Proof: $\mathcal{P}_0$ is nonempty. □

At this point it is moreover possible to prove some extra properties of S_0.

Proposition 4: negation for any proposition. There exists a continuous, strictly decreasing function $S_1 : [\mathcal{F}, \mathcal{T}] \to [\mathcal{F}, \mathcal{T}]$, such that $S_1(A|X) = S_0(A|X) = (\neg A|X)$ for all A and consistent X.

Proof: Let $\mathcal{P}_1$ be the set of all possible plausibility values $(A|X)$, where X is consistent. Now restrict the domain of S_0 to P_1. If $x_1, x_2 \in \mathcal{P}_1$ and $S_0(x_1) = S_0(x_2) = y$, then $x_1 = S_0(y) = x_2$, which means that S_0 is one-to-one. From Proposition 2 we know that $\mathcal{P}_1$ is the range of S_0. Then, since $\mathcal{P}_1$ is a dense subset of $(\mathcal{F}, \mathcal{T})$ and since S_0 is nonincreasing,

this implies that S_0 is continuous, as any discontinuity would produce a gap in the range of S_0. Now define $S_1(x) = \lim_{y \to x} S_0(y)$, so that S_1 is the required function. □

Note that $\mathcal{P}_0$ does not include the values $\mathcal{F}$ and $\mathcal{T}$. $\mathcal{P}_0$ is designed to guarantee that some set of three propositions can be assigned to any chosen set of three plausibilities, which is an essential ingredient for proving the product rule in section 3.3.1. It is however trivial to find propositions with plausibility values $\mathcal{F}$ and $\mathcal{T}$, namely the contradiction and tautology respectively. For this reason $\mathcal{P}_0$ need not include them.

Strengthening Axiom 4 to require that $\mathcal{P}_0 = [\mathcal{F}, \mathcal{T}]$ allows for omission of the requirement that S_0 be nonincreasing, as shown by Paris (1994). This is one of many examples where one part of Axiom 4 may be weakened at the cost of strengthening another part: the final result will be the same whichever path is chosen, though some may find a specific form of Axiom 4 more natural than another.

Axiom 5: conjunction. There exists a continuous function $F : [\mathcal{F}, \mathcal{T}]^2 \to [\mathcal{F}, \mathcal{T}]$, strictly increasing in both arguments on $(\mathcal{F}, \mathcal{T}]^2$, such that $(A \wedge B|X) = F((A|B, X), (B|X))$ for any A, B and consistent X.

This axiom sets three requirements on F. Let us start by discussing the reasonableness of the first two.

If one's state of information changes so that either B or $A|B$ becomes infinitesimally more plausible, then the plausibility of $A \wedge B$ should also change infinitesimally; it would be quite unnatural for the latter to suddenly jump up or down. Thus, we require that F be continuous.

Now assume that one's state of information changes such that either B or $A|B$ becomes more plausible, while the other is left no less plausible. If $A \wedge B$ were to now become less plausible, that would be rather counterintuitive. This then requires F to be strictly increasing in both arguments.

We do note that the use of the half-open interval $(\mathcal{F}, \mathcal{T}]$ is unusual. While continuity - given this condition - guarantees that F will be strictly increasing on the closed interval $[\mathcal{F}, \mathcal{T}]$ as well, there seems to be no reason here for selecting the half-open interval over the closed or fully-open intervals. Indeed, either of the latter two may be preferable due to symmetry considerations. This curious choice, however, does not seem to pose any issues for the further proof.

Finally, to understand why F takes precisely the arguments $(A|B, X)$ and $(B|X)$, one could first consider the obvious candidates $(A \wedge B|X)$ might depend on. These are $(A|X)$, $(B|X)$, $(A|B, X)$, and $(B|A, X)$. There are a total of fifteen different ways to combine these four plausibilities: four using only one, six using two different ones, four using three, and one using all four. These numbers depend on the requirement that a particular ordering of the arguments in F may be chosen without loss of generality; for example, one may set the convention that arguments containing the greatest number of propositions come first, followed by an alphabetic ordering convention.

There exists and additional symmetry, namely that $A \wedge B \Leftrightarrow B \wedge A$; this reduces the number of combinations to nine (two options are lost in each of the single, double and triple valued combinations). An argument that further exhausts the possible combinations was provided by Tribus (1969), and is presented in the following.

1. $(A \wedge B|X) = F(A|X)$. Take A to be a tautology and B any atomic proposition. Now $(A|X) = \mathcal{T}$, so $F(\mathcal{T}) = (B|X)$. But X can be chosen to make $(B|X)$ any value in $\mathcal{P}_0$, since different states of information X change the plausibility of B. Thus we

have arrived at a contradiction.

2. $(A \wedge B|X) = F(A|B, X)$. Let A and B to be the same atomic proposition. Then $(A \wedge B|X) = (A|X)$ and $(A|B, X) = \mathcal{T}$. Thus $F(\mathcal{T}) = (A|X)$, a contradiction by the same argument as before.

3. $(A \wedge B|X) = F((A|X), (A|B, X))$. Take A to be a tautology and B any atomic proposition. Now $F(\mathcal{T}, \mathcal{T}) = (B|X)$, yielding a contradiction yet again.

4. $(A \wedge B|X) = F((A|B, X), (B|A, X))$. Let A and B be the same atomic proposition. Now $(A|B, X) = (B|A, X) = \mathcal{T}$, so that $F(\mathcal{T}, \mathcal{T}) = (A|X)$, another contradiction by the same argument.

5. $(A \wedge B|X) = F((A|X), (B|X))$. Take A to be any atomic proposition and B to be $\neg A$. Now $(A \wedge B|X) = \mathcal{F}$, and we obtain $\mathcal{F} = F((A|X), S_1(A|X)) = F(x, S_1(x))$ for all $x \in \mathcal{P}_0$ by setting $x = (A|X)$. Alternatively, by setting A and B to be the same atomic proposition, we obtain $F(x, x) = x$. These two equalities lead to three undesirable consequences that will require the rejection of this option for F.

   - Assume that F is continuous, then $F(x, S_1(x)) = \mathcal{F}$ and $F(x, x) = x$ for all $x \in [\mathcal{F}, \mathcal{T}]$. S_1 is continuous and strictly decreasing, and has domain and range $[\mathcal{F}, \mathcal{T}]$. Hence it must have a fixed point $\mathcal{F} < x_0 < \mathcal{T}$ in this interval such that $S_1(x_0) = x_0$. We conclude that $x_0 = F(x_0, x_0) = F(x_0, S_1(x_0)) = \mathcal{F}$; a contradiction with the assumption that $x_0 > \mathcal{F}$. This means F is discontinuous, which is at odds with the assumptions of Axiom 5.

   - Set $x = (A|X) = (\neg A|X)$. Now $x = F(x, x) = F(x, S_1(x)) = \mathcal{F}$, which means $x = S_1(x) = \mathcal{F}$. However, this leads to $S_1(\mathcal{F}) = \mathcal{F}$, which cannot be possible since S_1 is strictly decreasing in the range $[\mathcal{F}, \mathcal{T}]$. Thus we are forced to conclude that a proposition and its negation cannot be equally plausible, which is very unnatural indeed.

   - Choose $x \in \mathcal{P}_0$ and define $y = S_1(x)$. We have defined $S_1(\mathcal{T}) = \mathcal{F}$ and thus obtain $y > \mathcal{F}$ since S_1 is strictly decreasing. Now we conclude that $F(x, y) = F(x, S_1(x)) = \mathcal{F}$, which is undesirable because it means that we can choose $x = (A|X) > \mathcal{F}$ and $y = (B|X) > \mathcal{F}$ such that $(A|X)$ and $(B|X)$ are individually plausible, with their conjunction $(A \wedge B|X)$ known to be false.

This leaves four remaining options:

1. $(A \wedge B|X) = F((A|B, X), (B|X))$
2. $(A \wedge B|X) = F((A|B, X), (A|X), (B|X))$
3. $(A \wedge B|X) = F((A|B, X), (B|A, X), (B|X))$
4. $(A \wedge B|X) = F((A|B, X), (B|A, X), (A|X), (B|X))$

Here Tribus shows that adopting 3 or 4 leads to situations where dealing with some inconsistent state of information Y cannot be avoided. This leads to ambiguities where a proposition and its negation are both true, which is clearly a problem for consistent states of information. However, it is well know that anything can be proven from inconsistent premises (Daepp and Gorkin 2003), so there is no problem with picking an arbitrary value for $(A|Y)$ when Y is inconsistent; indeed, this may be a strong reason to argue that $(A|Y) = \mathcal{T}$ in such a case. Tribus's argument here is therefore not especially convincing, according to Van Horn.

Finally, Tribus leaves ruling out 2 as an exercise to the student. At this point then, there

seems to be no absolutely compelling reason to choose one option over the other three. Option one does seem to be the most intuitively appealing, and has not garnered much controversy (Horn 2003). Furthermore, options two, three, and four simply add arguments to those already present in option one. The simplest option is generally preferred ceteris paribus, by application of Occam's Razor.

The following is an intuitive rationale for option one, provided by Jaynes (2003, Chapter 2). Here Jaynes's notation has been replaced by that of Van Horn (2003).

> *"In order for $A \wedge B$ to be a true proposition, it is necessary that B is true. Thus the plausibility $(B|X)$ should be involved. In addition, if B is true, it is further necessary that A should be true; so the plausibility of $(A|B, X)$ is also needed. But if B is false, then of course $A \wedge B$ is false independently of whatever one knows about A, as expressed by $(A|\neg B, X)$; if the robot reasons first about B, then the plausibility of A will be relevant only if B is true. Thus, if the robot has $(B|X)$ and $(A|B, X)$ it will not need $(A|X)$. That would tell it nothing about $A \wedge B$ that it did not have already."*

Terenin and Draper (2015) remark that many - including they themselves, Tribus (1969) and Fine (1973) - find this argument compelling. It thus seems reasonable to accept Axiom 5 at this point.

Just as was the case with S_0, Cox's original paper does not explicitly require F to be strictly increasing, but rather demands that it be twice differentiable (Cox 1946). Yet again however, careful reading of his proof shows that F is implicitly assumed to be either strictly increasing or decreasing in both arguments on $(\mathcal{F}, \mathcal{T}]$ (Horn 2003).

Arnborg and Sjödin (1999) have shown that requirements on F may be slightly relaxed, but the requirement that it be strictly increasing is in any case necessary. In particular, only requiring that F be nondecreasing opens the possibility that $F(x, y) = \min(x, y)$. This is in accord with Cox's requirements (Snow 1998), but does not lead to a system equivalent to probability theory.

It is worth mentioning that another system of axioms by Jaynes (2003) will lead to the same result (specifically Theorem 13 in section 3.3.3). He adopts the metaphor of a robot he wishes to program to reason and behave sensibly in the presence of uncertainty. This leads him to posit a number of 'common sense' desiderata, which take the place of our axioms. These desiderata are stated in natural language, which has the advantage that the assumptions he makes seem more intuitive than those Van Horn has selected. On the other hand, this lack of formality forces Jaynes to appeal to 'common sense' in a number of steps during his derivation. While not to the proof's detriment per se, some might find it less convincing than the more axiomatised treatment that is followed here.

3.3 Obtaining the rules of probability theory

The five axioms from the previous section and their immediate consequences are all that shall be required to prove the main result; namely that these assumptions lead to a system isomorphic with probability theory. This section is dedicated to that endeavour. In particular, we shall find that our plausibilities must obey the product and sum rules of probability theory. Section 3.3.1 attempts to prove the former, while section 3.3.2 is dedicated to the latter. Along this course we shall find that plausibilities may be restricted to the interval $[0, 1]$, thus completing the derivation. The conclusion of this chapter is summarised in Theorem 13 of section 3.3.3.

3.3.1 The product rule

Let us start by taking a closer look at F, the function that characterises conjunction between plausibilities.

Proposition 5: associativity of F. $F(x, F(y, z)) = F(F(x, y), z))$ for all $x, y, z \in [\mathcal{F}, \mathcal{T}]$.

Proof: For consistent X, apply twice Axiom 5 and once 2.3 to the expression $(A \wedge B \wedge C|X)$.

$$\begin{aligned}(A \wedge B \wedge C|X) &= F\left[(A|(B \wedge C), X), (B \wedge C|X)\right] \\ &= F\left[(A|(B \wedge C), X), F\left[(B|C, X), (C|X)\right]\right] \\ &= F\left[(A|B, C, X), F\left[(B|C, X), (C|X)\right]\right]. \end{aligned} \tag{3.1}$$

And alternatively

$$\begin{aligned}(A \wedge B \wedge C|X) &= F\left[(A \wedge B|C, X), (C|X)\right] \\ &= F\left[F\left(A|B, C, X\right), (B|C, X)\right], (C|X)\right]. \end{aligned} \tag{3.2}$$

Equation these two expressions yields

$$F\left[(A|B, C, X), F\left[(B|C, X), (C|X)\right]\right] = F\left[F\left(A|B, C, X\right), (B|C, X)\right], (C|X)\right]. \tag{3.3}$$

Now invoking Axiom 4 we may set $(A|B, C, X) = x$, $(B|C, X) = y$ and $(C|X) = z$ for any $x, y, z \in \mathcal{P}_0$ to obtain

$$F\left[x, F\left[y, z\right]\right] = F\left[F\left[x, y\right], z\right]. \tag{3.4}$$

Note that this is the reason we required Axiom 4 to handle cases of at least three separate propositions. By assumption $\mathcal{P}_0$ is dense on $(\mathcal{F}, \mathcal{T})$ and F is continuous in this same interval. As such the equation holds for all $x, y, z \in [\mathcal{F}, \mathcal{T}]$. □

This equation may be recognised by those expressly familiar with functional equations. Aczél (1966) calls it 'the associativity equation' and as we shall see it forces plausibilities x, y and z to obey the product rule of probability theory.

Aczél derives the general solution to (3.4) without assuming differentiability; unfortunately, his proof spans 11 pages. Here a shorter proof by Cox (1961) that assumes differentiability is provided, as edited by Jaynes (2003). Specifically, the following proposition shall be proven:

Proposition 6: product rule for proto-probabilities w. There exists a continuous, strictly increasing or decreasing function w such that $w(A, B|X) = w(A|B, X)w(B|X)$ for any propositions A, B and for consistent X.

Proof: The associativity equation (3.4) has the trivial solution $F(x, y) =$ constant, but this does not satisfy the requirement that F be strictly increasing (Axiom 5). Furthermore, this solution does not allow us to draw any reasonable conclusions from the conjunction of propositions, so it is useless for the purposes of Cox. Note that the Cox-Jaynes approach

would fail here if no more general solution could be found. Fortunately, such a solution does exist.

Let us abbreviate $u = F(x, y)$ and $v = F(y, z)$. Equation (3.4) reduces to

$$F(x, v) = F(u, z). \tag{3.5}$$

Both sides can be differentiated with respect to x and y. F_i is used to denote the derivative of F to its i'th argument. This yields two equations

$$F_1(x, v) = F_1(u, z)F_1(x, y), \tag{3.6}$$
$$F_2(x, v)F_1(y, z) = F_1(u, z)F_2(x, y). \tag{3.7}$$

From (3.6) we obtain $F_1(u, z) = F_1(x, v)/F_1(x, y)$ and plugging this into equation (3.7) gives

$$G(x, v)F_1(y, z) = G(x, y), \tag{3.8}$$

where $G(x, y) = F_2(x, y)/F_1(x, y)$. This shows that $G(x, v)F_1(y, z)$ is independent of z. Multiplying (3.8) by $F_2(y, z)/F_1(y, z) = G(y, z)$ yields

$$G(x, v)F_2(y, z) = G(x, y)G(y, z). \tag{3.9}$$

Denoting the left-hand sides of (3.8) and (3.9) by U and V respectively, we note that $0 = \partial U/\partial z = \partial V/\partial y$, since U is independent of z. Thus we must conclude that $V = G(x, y)G(y, z)$ is independent of y. The most general function $G(x, y)$ that satisfies this requirement is

$$G(x, y) = r\,\frac{H(x)}{H(y)}. \tag{3.10}$$

Where r is a constant and H an arbitrary function. This way, $G(x, y)$ can still depend on y, while $G(x, y)G(y, z)$ does not. We have $G > 0$ since F is strictly increasing in both arguments, so $r > 0$ and $H(x)$ may not flip sign for $x \in [\mathcal{F}, \mathcal{T}]$. Plugging (3.10) into (3.8) and (3.9) gives respectively

$$F_1(y, z) = \frac{H(v)}{H(y)}, \tag{3.11}$$

$$F_2(y, z) = r\,\frac{H(v)}{H(z)}. \tag{3.12}$$

Taking a step back, we write $dv = dF(y, z) = F_1 dy + F_2 dz$ which can be rewritten as

$$\frac{dv}{H(v)} = \frac{dy}{H(y)} + r\,\frac{dz}{H(z)}. \tag{3.13}$$

And it follows that

$$\exp\left\{\frac{dv}{H(v)}\right\} = \exp\left\{\frac{dy}{H(y)}\right\}\exp\left\{r\frac{dz}{H(z)}\right\}. \tag{3.14}$$

We note here that Jaynes integrates (3.13) over v before taking the exponent, presumably so as to not encounter possible issues with infinitesimals later. We have opted to skip this step, as the purpose of this proof is for it to be illustrative, and this decision simplifies our definitions in the following. We again refer to Aczél (1966) for a fully closed proof.

Continuing by defining

$$w(x) \equiv \exp\left\{\frac{dx}{H(x)}\right\}, \tag{3.15}$$

we obtain

$$w\left[F(y,z)\right] = w(v) = w(y)w^r(z). \tag{3.16}$$

Applying (3.16) to equation (3.5) gives

$$w(x)w^r(v) = w(u)w^r(z). \tag{3.17}$$

Finally, applying (3.16) yet again yields

$$w(x)w^r(y)w^{r^2}(z) = w(x)w^r(y)w^r(z), \tag{3.18}$$

so we must have $r = 1$. Then from (3.16) and the definition for y and z we find

$$w(B, C|X) = w(B|C, X)\, w(C|X), \tag{3.19}$$

for any propositions B, C and consistent X. The identification $B \to A$ and $C \to B$ completes the proof. □

This looks already much like the familiar product rule for probabilities. Of course, at this point w may not yet be identified with probabilities, but - using nomenclature by Van Horn (2003) - they can be thought of as 'proto-probabilities' that rescale the plausibilities and satisfy the product rule for probabilities.

Proposition 7: identification of $w(\mathcal{F})$ and $w(\mathcal{T})$. $w(\mathcal{F}) = 0$, $w(\mathcal{T}) = 1$ and $0 < w(x) < 1$ for $\mathcal{F} < x < \mathcal{T}$.

Proof: By construction (equation (3.15)) $w(x)$ is positive, continuous and strictly increasing or strictly decreasing, depending on the sign of $H(x)$. Now consider the proposition $(A \wedge D|X)$, where A is any proposition and D is a tautology for consistent X. Then

$$w(A|X) = w(A \wedge D|X) = w(D|A, X)w(A|X) = w(\mathcal{T})w(A|X). \tag{3.20}$$

Since A is any proposition, it must be true that $w(\mathcal{T}) = 1$. Next, consider the case where $(A|X) = \mathcal{F}$ and B is any proposition for consistent X. Since $(A|X)$ is impossible, $(A \wedge B|X)$ must also be impossible, thus $(A|X) = (A \wedge B|X)$. Furthermore, if A is impossible given the state of information X, then adding extra information B does not make A any more possible, hence $(A|B, X) = (A|X)$. Plug this into equation (3.19) to obtain

$$\begin{aligned} w(A \wedge B|X) &= w(A|B,X)w(B|X) \\ w(A|X) &= w(A|X)w(B|X) \\ w(\mathcal{F}) &= w(\mathcal{F})w(B|X), \end{aligned} \tag{3.21}$$

and this must be true for any proposition D. This property is only satisfied for $w(A|X) = 0$ or $w(A|X) = \pm\infty$. Note that $w(A|X) = -\infty$ is ruled out by the condition that w be positive. The options $w(x) = 0$ and $w'(x) = +\infty$ are equivalent in content: since w is continuous and either strictly increasing or decreasing, both cases can be mapped onto each other by the simple transformation $w'(x) = 1/w(x)$. We can thus without loss of information adopt the convention that $0 < w(x) < 1$ for $\mathcal{F} < x < \mathcal{T}$. □

This proposition shows that the proto-probabilities share the same range of values as true probabilities, and represent truth and falsity the same way. The next - and final - step in Cox's derivation is to show that these proto-probabilities can be mapped to satisfy the sum rule of probability theory. The following section does exactly that.

3.3.2 The sum rule

Under the mapping from plausibilities to proto-probabilities it has become apparent that F essentially contains the same information as the product rule of probability theory. Similarly, one might expect the remaining function S_1 to have a correspondence with the sum rule. This is indeed what will be shown below.

Definition 9: Define a new function $S(x)$ to be $p(S_1(p^{-1}(x)))$, where

- $p(x) = w^r(x)$
- $r = -\log 2/\log w(\alpha)$
- α is the unique fixed point of S_1, i.e. $S_1(\alpha) = \alpha$ and $\mathcal{F} < \alpha < \mathcal{T}$.

The purpose of this section is to show that the function p satisfies the sum rule of probability theory. Note that it satisfies the product rule as well, since it is a power of w. It will in the following become clear that p is the plausibility function we have been searching for.

Proposition 8: properties of p and S. S and p^{-1} are well defined. Furthermore, $S(1/2) = 1/2$ and lastly p has the following properties:

1. p is continuous and strictly increasing.
2. $p(\mathcal{F}) = 0$, $p(\mathcal{T}) = 1$ and $0 < p(x) < 1$ for $\mathcal{F} < x < \mathcal{T}$
3. $p(A \wedge B|X) = p(A|B,X)p(B|X)$ for all A, B and consistent X.

Proof: S_1 has a fixed point because it is continuous and maps $[\mathcal{F}, \mathcal{T}]$ onto itself. The fixed point must be unique since S_1 is strictly decreasing, which means α is well defined and $\mathcal{F} < \alpha < \mathcal{T}$. Then, by Proposition 6, $0 < w(\alpha) < 1$ from which it follows that $-\infty < \log w(\alpha) < 0$. Moreover, this implies $0 < r < \infty$ by Definition 9. Having taken w to be strictly increasing and continuous, we see that r is strictly increasing and continuous in α, which proves property 1 via Definition 5. Property 1 means p is bijective, so that its inverse p^{-1} is well defined. With $r > 0$ the second item is also proven, since w has those same properties. The last item follows because for all A, B and consistent X we have

$$p(A \wedge B|X) = w(A \wedge B|X)^r = w(A|B,X)^r w(B|X)^r = p(A|B,X)p(B|X), \tag{3.22}$$

since $(ab)^r = a^r b^r$. Note that this means p satisfies the product rule of probability theory. Rewriting the definition of r yields

$$w(\alpha) = 2^{-\log 2/r} = \frac{1}{2}^{1/r}, \tag{3.23}$$

so that $p(\alpha) = w(\alpha)^r = \frac{1}{2}$. Finally then

$$S\left(1/2\right) = p(S_1(p^{-1}\left(1/2\right))) = p(S_1(\alpha)) = p(\alpha) = 1/2. \tag{3.24}$$

□

At this point we summarise Propositions 2 and 4 together with Axiom 4 in terms of p and S.

Proposition 9: further structure in p and S. p and S have the following properties:

1. $p(\neg A|X) = S(p(A|X))$ for all A and consistent X.
2. S is continuous and strictly decreasing.
3. $S(S(x)) = x$ for all $x \in [0,1]$.
4. Define $\mathcal{P} = p(\mathcal{P}_0)$. Then $\mathcal{P}$ is a dense subset of $(0,1)$. Furthermore, for every $y_1, y_2, y_3 \in \mathcal{P}$ there exists some consistent X with a basis of three atomic propositions A_1, A_2, A_3, such that $p(A_1|X) = y_1$, $p(A_2|A_1,X) = y_2$ and $p(A_3|A_2,A_1,X) = y_3$.

Proof: By the definition of S we have: $S(p(A|X)) = p(S_1(p^{-1}(A|X))) = p(S_1(A|X)) = p(\neg A|X)$, which proves point 1. Point 2 is true by the definition of S together with the fact that p and S are both continuous, with p strictly increasing and S strictly decreasing. As for the third point, by Propositions 2 and 4:

$$S(S(x)) = p(S_1(p^{-1}(p(S_1(p^{-1}(x)))))) = p(S_1(S_1(p^{-1}(x)))) = p(p^{-1}(x)) = x. \tag{3.25}$$

Item 4 follows from Axiom 4 and continuity of p. □

The Associativity Equation (3.4) for F was imperative for the derivation of the product rule of probability theory. Similarly, a function equation for S is required to enable a derivation of the sum rule for p. This is the content of the following proposition:

Proposition 10: a functional equation for S. For all $0 < x \leq y < 1$ it is true that $yS(x/y) = S(x)S\left(\frac{S(y)}{S(x)}\right)$.

Proof: For any $u, y \in \mathcal{P}$ choose propositions A, B and a consistent X such that $y = p(B|X)$ and $u = p(A|B,X)$. Let $x = uy = p(A \wedge B|X)$ and note that $u, x, y \in (0,1)$. Now write

$$yS(x/y) = yS(u) = p(B|X)p(\neg A|B,X) = p(\neg A \wedge B|X), \tag{3.26}$$

and define $C \equiv (\neg A \vee \neg B)$, $D \equiv (A \vee \neg B)$. Note that

$$C = (\neg A \vee \neg B) = \neg(A \wedge B), \tag{3.27}$$
$$C \wedge D = (\neg A \vee \neg B) \wedge (A \vee \neg B) = \neg B, \tag{3.28}$$
$$\tag{3.29}$$

and thus

$$S(y) = p(\neg B|X) = p(C \wedge D|X), \tag{3.30}$$
$$S(x) = p(A \wedge B|X) = p(C|X). \tag{3.31}$$

Then, rewriting $C \wedge \neg D$ as

$$C \wedge \neg D = (\neg A \vee \neg B) \wedge \neg(A \vee \neg B) = (\neg A \vee \neg B) \wedge (\neg A \wedge B) = (\neg A \wedge B) \tag{3.32}$$

allows us to write

$$\begin{aligned} S(x)S\left(\frac{S(y)}{S(x)}\right) &= S(x)S\left(\frac{p(C \wedge D|X)}{p(C|X)}\right) \\ &= p(C|X)S\left(p(D|C,X)\right) \\ &= p(C|X)p(\neg D|C,X) \\ &= p(C \wedge \neg D|X) \\ &= p(\neg A \wedge B|X) = yS(x/y). \end{aligned} \tag{3.33}$$

Note that a state of information C, X was used in this derivation. This state is consistent by Axiom 2.4 since $p(\neg C|X) = x < 1$. Equation (3.33) holds for all $0 < y < 1$, $0 < u \leq 1$ since $\mathcal{P}$ is dense and S is continuous. Therefore it also holds for all $0 < x \leq y < 1$. □

At this point, Van Horn refers to Paris (1994) for a proof of the following proposition (Horn 2003, Lemma 12). We shall do so as well, for the simple reason that the full proof is rather elaborate. Jaynes (2003, Chapter 2.2) provides an alternative - four page long - proof of the sum rule as well.

Proposition 11: negation. Let $f : [0,1] \to [0,1]$ be a strictly decreasing and continuous function with $f(0) = 1$, $f(1) = 0$ and $f(1/2) = 1/2$. If f satisfies the following two functional equations for all $0 < x \leq y < 1$:

- $f(f(x)) = x$
- $y\, f(x/y) = f(x)\, f(f(y)/f(x))$,

then $f(x) = 1 - x$ for all $0 \leq x \leq 1$.

Proof: We refer to lemmas 3.10 to 3.15 of Paris (1994, p.29-32) for a proof of this proposition. □

Finally, we derive the sum rule of probability theory.

Proposition 12: the sum rule. $p(\neg A|X) = 1 - p(A|X)$ for all A and consistent X.

Proof: By Propositions 8, 9.3 and 10, S satisfies the conditions of Proposition 11, and hence $S(x) = 1 - x$ for all $0 \leq x \leq 1$. Then by Proposition 9.1 we have $p(\neg A|X) = S(p(A|X)) = 1 - p(A|X)$ for all A and consistent X. □

Thus p satisfies both the sum and product rules of probability theory. Moreover it maps propositions to the range $[0, 1]$ up to some transformation that leaves the content of p invariant. It is now clear that probability theory is the unique system of plausible reasoning that Cox was searching for.

The following section summarises these results, and derives a more general sum rule that may be more familiar to some.

3.3.3 Probability theory

In the previous it has been shown that p satisfies the familiar product and sum rules. These are exactly the basic rules of probability theory, and as such p can be thought of as a mapping from Cox's system of plausible reasoning to probability theory. Note that no information is lost in this mapping, since p is invertible. It has thus been shown that Cox's system is isomorphic to probability theory. The properties of p may be summarised in one final theorem.

Theorem 13: probability theory. There exists a continuous, strictly increasing function p such that for all A, B and consistent X,

1. $0 \leq p(A|X) \leq 1$
2. $p(A|X) = 0 \Leftrightarrow A$ is known to be false given the state of information X.
3. $p(A|X) = 1 \Leftrightarrow A$ is known to be true given the state of information X.
4. $p(A \wedge B|X) = p(A|B, X)p(B|X)$
5. $p(\neg A|X) = 1 - p(A|X)$

Given that these are all the basic rules, one might now wonder how to derive the probability of a statement such as $p(A \vee B|X)$, since it appears a rule for calculating this is lacking from Theorem 13. Jaynes (2003, Chapter 1.6) shows however that conjunction (Theorem 13.4) and negation (Theorem 13.5) is an 'adequate set of operations'; operations from which all logic functions can be constructed. The derivation of $p(A \vee B|X)$ from the basic rules shall serve as a demonstration of this concept.

$$\begin{aligned} p(A \vee B|X) &= p(\neg(\neg A \wedge \neg B)|X) \\ &= 1 - p(\neg A \wedge \neg B|X) \\ &= 1 - p(\neg A|\neg B, X)p(\neg B|X) \\ &= 1 - (1 - p(A|\neg B, X))\ p(\neg B|X) \\ &= 1 - p(\neg B|X) + p(A \wedge \neg B|X) \\ &= p(B|X) + p(A \wedge \neg B|X) \\ &= p(B|X) + p(\neg B|A, X)p(A|X) \\ &= p(B|X) + (1 - p(B|A, X))\, p(A|X) \\ &= p(B|X) + p(A|X) - p(A \wedge B|X). \end{aligned} \tag{3.34}$$

Note that if $B = \neg A$, this simplifies to

$$1 = p(A \vee \neg A|X) = p(\neg A|X) + p(A|X) - p(A \wedge \neg A|X) = p(\neg A|X) + p(A|X), \quad (3.35)$$

which is exactly the sum rule. Thus, the generalised sum rule equation (3.34) was derived through successive applications of the basic sum and product rules. This illustrates the basic concept of an adequate set of operations.

A final consideration is that of consistency of the system. The question of whether there actually exists a choice for $\mathcal{F}, \mathcal{T}, F, S_0, \mathcal{P}_0, (\cdot|\cdot)$ and a definition of a state of information that satisfies Axioms $1-5$ is an important one, because Cox's theorem has no useful content without such consistency. Van Horn (2003) argues that Kolmogorov's set-theoretical approach to probability theory (Kolmogorov 1933) qualifies as an existence proof for our system. Jaynes (2003) derives an equivalent of Theorem 13 by different means, which may also be interpreted as an indication that our system is consistent.

Terenin and Draper (2015) provide a further rigorisation of the CJ approach, and in the process find that the Kolmogorov axioms are a special case of the CJ system; i.e. under their approach Kolmogorov's axioms become theorems. They furthermore provide a theorem by Stone (Stone 1936) that shows an equivalence between the two systems, strengthening Van Horn's claim that the Kolmogorov axioms prove the existence of a consistent CJ system. It thus seems reasonable to accept these claims of consistency.

This has been an overview of the Cox's theorem. We have followed the approach of Van Horn (2003), although we have included elements of Jaynes (2003) and Paris (1994). as well. Any theorem is only as strong as its assumptions however, and the CJ approach faces a number of challenges to its axioms. For this reason, the next chapter is dedicated to exploring the controversies surrounding Cox's theorem.

Chapter 4

Controversy surrounding the Cox-Jaynes approach

The Cox-Jaynes approach is not without controversy. While it is generally agreed that the conclusions (that is, Theorem 13) are valid given the assumptions, the validity of the five axioms has often been questioned. Note that even though we refer to our proof as the CJ approach, neither Cox nor Jaynes provided this exact derivation. Our proof is based mainly on Van Horn (2003), though it incorporates elements of Jaynes (2003) and Paris (1994). Jaynes based his own proof on 'common sense' axioms, some of which have an equivalent in our axioms; for others the exact relation is less evident. Cox's proof is more in line with that provided here, although he originally omitted explicitly mentioning Axiom 4. The importance of this will be discussed below.

The reason for providing a different proof than Cox and Jaynes is that this variant is more explicit in its exact assumptions. This allows for a more stark discussion of the controversy surrounding it. Indeed, Van Horn himself mentions that his paper's goal is to "(...) understand the significance, applicability, and limitations of Cox's theorem (...)", which prompted him to follow Paris's lead in *"(...) carefully and explicitly laying out the requirements we use."*

Taking advantage of the clarity provided by Van Horn's proof, we can now set forth the three main points of controversy with the CJ approach.

The main points of controversy:

1. The assumption that plausibilities can be represented by real numbers (Axiom 1) will be examined in section 4.1.

2. The density assumption (Axiom 4), which ties into the question of proposition domain size, is discussed in section 4.2.

3. The necessity of countable additivity of the probability function p is considered in section 4.3.

The first two have been the subject of discussion throughout the literature, while the third is only explicitly mentioned by Terenin and Draper (2015) and may in a sense be thought of as an extension of the second issue.

4.1 Plausibilities as real numbers

Here we shall discuss the assumption that plausibilities can be represented by real numbers, given by Axiom 1 as follows:

Axiom 1: representation. The plausibility of A given X, $(A|X)$ is a real number. There furthermore exists a number $\mathcal{T}$ such that $(A|X) \leq \mathcal{T}$ for all A and all X.

There have been a number of arguments against the reasonableness of this axiom. Many focus on the fact that the axiom assumes universal comparability between propositions: the idea that the plausibility of any proposition can be compared with that of every other proposition. Some find this assumption unwarranted, and hence we shall first discuss this criticism in section 4.1.1. As we shall see, for many the position that universal comparability is unwarranted is a consequence of the concern that one-dimensional theories - which automatically obtain universal comparability - cannot adequately represent ignorance. This is a special case of the concern that many critics of Bayesian probability theory have regarding the subjectivity of priors; namely, how to find priors that objectively represent one's knowledge. This specific concern will be examined in section 4.1.2. Others accept universal comparability, but see no need for assigning real numbers to propositions; a view that will be discussed in section 4.1.3.

4.1.1 Universal comparability

The reasonableness of universal comparability is - perhaps somewhat ironically - not universally accepted. The view that it is an unwarranted assumption arises from either of two beliefs, according to Jaynes (2003, Appendix A.4). The first belief is that human brains cannot compare all propositions, which means there is no reason for a theory of plausible reasoning to do so either. The second is the belief that a counterexample to the validity of universal comparability has been produced. Jaynes then argues that the first point is irrelevant, arguing that *"human brains do many absurd things while failing to do many sensible things."* Indeed, the purpose of developing a formal theory of inference is in part precisely to correct such reasoning.

The second point, he argues, is based on a misunderstanding of the theory. He claims that the counterexamples are all of the type where one tries to classify propositions based on multiple attributes, though perhaps notably provides no reference for this. An example of such an argument is that a mineralogist classifying rocks on both density (d) and hardness (h), will run into trouble ordering rocks when those two attributes vary somewhat independently. A solution might be to determine a function $f(d, h)$ that tells the mineralogist how to trade a change in d for a change in h and vice-versa, such that they will again be classifying on the basis of one attribute, namely f.

Since the developed theory by definition classifies proposition according to only one attribute - intuitively called 'degree of plausibility' - these counterexamples provide no indication that universal comparability would be fundamentally impossible. Jaynes then argues that, with this understanding, the possibility of representation by real numbers need not be questioned, while the desirability of doing so *"is attested to by all the nice results and useful applications of the theory"*.

It is of note that Jaynes remarks on the possible usefulness of multi-dimensional theories of plausibility for modelling human minds (2003, Chapter 1.8). He notes that for humans many propositions invoke additional (emotional) judgments beyond just that of plausibility; such a judgment would require many more than one coordinates to describe.

He provides the rather astute example that, while a proposition such as 'The refractive index of water is less than 1.3' would also in human brains not invoke much judgment, the proposition 'Your mother-in-law just wrecked your new car' generates a state of mind with many coordinates.

It thus seems that the first objection is indeed rather weak, while the second may be more valid if a stronger counterexample has been produced. The author has however been unable to locate such a counterexample, and it thus seems that this examination is in favour of universal comparability for now. A perhaps more interesting objection to universal comparability is the existence of two-dimensional theories of plausibility. This objection is examined in the following section.

4.1.2 Two-dimensional theories of plausibility

Two-dimensional theories of plausibility use two seperate numbers to describe one's belief in a proposition. It may seem unintuitive to do so, since a single real has been sufficient for our purposes. However, one motivation for preferring two-dimensional theories is the concern that ignorance cannot be adequately represented by a one-dimensional theory. Describing plausibility with a pair of numbers allows for a description of ignorance that cannot be achieved by a single real number.

Two well known two-dimensional theories are Dempster-Shafer belief-functions (Shafer 1976; Smets 1990) and possibility theory (Dubois and Prade 1988). Both of these systems represent one's certainty in a proposition by two numbers, motivated by the wish to separate belief in and ignorance of a proposition. Regarding belief-functions, Shafer (1976, p.42) writes that one's belief about a proposition cannot be fully described by one's degree of belief alone, since that does not include information about how much one doubts the proposition. Dubois and Prade (1988, p.11) write similarly of possibility theory that there is only a weak link between the possibility of one event and its contrary event. A second number they call 'necessity' is required to fully describe the event's uncertainty.

It is clear that such descriptions contain more information than those in a one-dimensional theory. Indeed, it may be for this reason that two-dimensional theories have been gaining momentum since the 1990s, when its comprehensive foundations were laid due to Walley (2000), Kuznetsov (1991) and Weichselberger (2000).

Currently, an increasingly popular two-dimensional theory goes by the moniker of 'imprecise probability theory' (Coolen et al. 2010). This is a generalisation of probability theory that utilises intervals of probability, instead of precise probability values. Under imprecise probability theory one may assign a wide interval of probability to a proposition in order to convey uncertainty about the exact probability. Such an assignment may be useful in situations where little information is available to determine a more exact plausibility. A brief discussion of this theory can be found in Appendix B.2.

The two-dimensional theories mentioned here are not excluded by Jaynes's mineralogist-argument from the previous section, since they're based on the view that the attribute 'degree of plausibility' it not sufficient for a full description. Such theories then necessarily lack universal comparability.

At this point, one may wonder how one-dimensional Bayesian theories represent ignorance. Historically, common practice for this has been using Laplace's principle of indifference: assigning equal probabilities to every proposition if these propositions are mutually exclusive.

There are various reasons to doubt the validity of this principle however; the most pragmatic one may be that it's unclear from the principle alone about exactly what we should be indifferent. Adopting the specific example of a parameter estimation problem, one may note that the principle of indifference does not express in which parameters indifference should hold. When estimating the variance σ^2 of a normal distribution $N(0, \sigma^2)$ it is not clear whether the prior should be uniform over σ, σ^2, or any other power of sigma.

Indeed, that Laplace's principle was so narrow is one of the reasons that frequentist methods replaced Bayesian methods at the start of the twentieth century (Gigerenzer et al. 1989). Even Cox himself (1961, p.29-34) argues against the principle of indifference on the basis of a contradiction in Boolean algebra (section B.1). Thus, while Cox's theorem proves probability theory to be the only proper one-dimensional description of plausibility, Cox also indicates that a one-dimensional theory may not be sufficient for all purposes.

More rigorous methods thus seem required to determine when indifference can be used properly, and what other priors should be used if it cannot. Priors constructed by such rigorous methods are said to be objective, as they are decided by mathematical rules, rather than personal belief. Harold Jeffreys was the first to rigorise ideas on constructing objective priors representing ignorance (Jeffreys 1939). The famous Jeffreys's rule is one of his accomplishments, and satisfies some invariance properties that the uniform prior lacks (Appendix A.1). However, although Jeffreys is considered to be an objective Bayesian, he did not hold the viewpoint that representation of ignorance should be unique in the sense that that there is only one logically correct prior. Instead, he relied on convention for consistency between problems, comparing the choice of initial prior to the choice of a system of units (Kass and Wasserman 1996).

Other methods of construction noninformative (i.e. ignorance) priors are maximum entropy methods (Jaynes 1957a; Jaynes 1957b; Jaynes 1963 and Appendix A.2), prior transformations (Jaynes 1968 and Appendix A.3), reference priors (Bernardo 1979; Berger and Bernardo 1991), and many more. We refer to Kass and Wasserman (1996) for a comprehensive overview. It may now be apparent that objective Bayesians are met with the difficulty that there are often too many available choices for constructing such priors to agree on a solution, which conflicts with its fundamental idea of a unique representation of reasonable belief. Indeed, the inability to properly deal with complete ignorance is perhaps the central criticism that objective Bayesians face.

It is clear by the above then, that one-dimensional theories are not without controversy. Van Horn, though, remarks on one final - and refreshingly pragmatic - reason for accepting one-dimensional theories, namely that they are the simplest available options. Axiom 1 therefore seems like a good property to have, provided it leads to a useful theory.

Finally, we note here that Axiom 3 essentially requires the one-dimensionality provided by Axiom 1: if $(A|X)$ and $(\neg A|X)$ were allowed to vary independently we would have constructed a two-dimensional theory. As such, the controversy surrounding Axiom 1 extends to Axiom 3 as well.

4.1.3 Qualitative ordering theories

Universal comparability alone, however, is not sufficient to support Axiom 1. We still require the additional assumption that plausibilities are represented by numbers at all. This section examines the validity of this assumption. In particular, we will consider the possibility of theories that obey universal comparability, but do so without explicitly assigning plausibilities in the form of (real) numbers.

Theories of this form certainly seem possible. For instance, one can satisfy universal comparability by admitting qualitative ordering of the form $(A|X) \geq (B|X)$ when A is deemed more plausible than B given a state of information X. Then, by offering no quantitative comparison beyond that, one can avoid assigning numbers as plausibilities and hence not satisfy Axiom 1. The work of Savage (1954) is an example of such a system. Fine (1973) has moreover provided a summary of attempts that do away with representation by real numbers.

Such methods may seem more natural than those that do satisfy Axiom 1, since they do not require a specific representation of plausibilities. However, Jaynes (2003, Appendix A.3) argues that these qualitatively comparative theories are still essentially equivalent to Axiom 1. He notes in particular that such systems will require transitive ordering: if $(A|X)$ is more plausible than $(B|X)$, and $(B|X)$ is more plausible than $(C|X)$, then $(A|X)$ must necessarily be more plausible than $(C|X)$, regardless of whether a number was assigned to those propositions. It must furthermore always be possible to extend the system by including more propositions, with their own ordering relations with regards to all other propositions; otherwise the method is needlessly narrow. These ordering relations must then also satisfy transitivity, and - by repeating this extension indefinitely - the set of propositions must in a sense become *"everywhere dense on the path from impossibility to certainty"*. Then those ordering relations may as well be replaced by real-numbered probabilities.

This has by no means been an exhaustive treatment of quantitative ordering theories. It does however seem plausible that those that satisfy transitive ordering essentially satisfy Axiom 1 as well. It furthermore seems reasonable that plausible reasoning systems should be able to compare plausibilities of all propositions with one another, thus necessitating transitive ordering of these propositions. Quantitative ordering theories must therefore not necessarily be at odds with Axiom 1.

Of the three objections raised in this section, it seems that only the problem of representing ignorance in one-dimensional theories remains an issue. Two-dimensional theories solve this problem, but necessarily lack universal comparability and thus do not satisfy Axiom 1. It remains unclear whether two-dimensional theories are strictly more useful than their one-dimensional counterpart; the latter at the very least constitutes a simpler description that nonetheless has access to various techniques for representing ignorance. At this point it seems reasonable to accept Axiom 1, while nevertheless acknowledging its limitations. At the very least, Cox's theorem proves that probability theory is the only proper one-dimensional representation of plausibility.

4.2 The density axiom

The second main point of controversy surrounding Cox's theorem focuses on what we have called the density axiom (Axiom 4):

Axiom 4: density. There exists a nonempty set of real number $\mathcal{P}_0$ with the properties that:

1. $\mathcal{P}_0$ is a dense subset of $(\mathcal{F}, \mathcal{T})$. That is, for every pair of real numbers a, b such that $\mathcal{F} \leq a < b \leq \mathcal{T}$, there exists some $c \in \mathcal{P}_0$ such that $a < c < b$.
2. For every $y_1, y_2, y_3 \in \mathcal{P}_0$ there exists some consistent X with a basis of at least three atomic propositions A_1, A_2, A_3, such that $(A_1|X) = y_1$, $(A_2|A_1, X) = y_2$ and $(A_3|A_2, A_1, X) = y_3$.

When this axiom was introduced we presented Van Horn's reasons for its validity, and in sections 3.3.1 and 3.3.2 it was integral for the derivation of the product and sum rules of probability. Of all of Cox's axioms it is also perhaps discussed the most in the literature. This is especially interesting since Axiom 4 did not appear explicitly in Cox's original work. It may be controversial for precisely that reason, since the form in which it was originally presented allowed for an interpretation that permitted the construction of a counterargument to Cox's theorem by Halpern (1996).

The circumstances that lead to this counterexample are elaborated on in section 4.2.1. The fixed version of the axiom that invalidates Halpern's counterexample is functionally equivalent to our Axiom 4, but the additional requirements bear their own controversy to do with the cardinality of Cox's plausibility function's domain. This will be discussed in section 4.2.2.

4.2.1 Cox's alleged omission

Axiom 4 is special in the sense that it did not appear explicitly in Cox's original work, and was only later supplied by Paris (1994), who noted its necessity. Cox himself relied on more sporadically introduced assumptions, which has lead to some confusion. In this section we examine this confusion briefly.

As Cox did not use Axiom 4 in his derivation, he required a number of additional assumptions for his proof. Notably, without the density axiom it is unclear for what values of x, y, z his equivalent of the associativity equation (3.4) should hold. Upon reading Cox's original paper it may reasonably be concluded that the equation is only considered for some numbers in a finite (i.e. not dense) probability distribution. This is exactly what Halpern (1996) did, and it allowed him to construct a counterexample to Cox's theorem by constructing a set of numbers that satisfied all of Cox's axioms, but did not yield an associative function F such as in (3.4). Snow (1998) argues that Cox did mention the caveat that equation (3.4) must hold for arbitrary x, y, z, thereby dismissing Halpern's counterexample. Indeed, on page 6 of his 1946 paper Cox writes

> *"The function F must be such as to satisfy Eq.* (8) *for arbitrary values of x, y, and z."*

where Eq. (8) refers to the equivalent of (3.4). Halpern's position however is understandable, since - as Snow mentions in a later work (2002) - Cox's three papers on this subject vary somewhat in their explicit assumptions and derivation (Cox 1946; Cox 1961; Cox 1978).

Implicitly, then, Cox took the domain of his plausibilities to be dense, as Axiom 4 proposes. Paris (1994) was the first to formalise this requirement in a form similar to the axiom that Van Horn (and we) adopted. It is of note that the adoption of a state of information has made Axiom 4 more intuitively appealing that the original formulation by Paris. This is remarked on by Terenin and Draper (2015), though they criticise the non-rigorousness of the state of information concept, and therefore find it difficult to evaluate its validity formally. Some slightly varying formulations of Paris's density axiom are discussed by Halpern (1996; 1999).

It thus seems that Halpern's concern does not apply to the proof of Cox's theorem as presented in section 3.3. However, the density axiom itself implies a consequence that some find unwarranted; namely a restriction of the proposition domain to infinite size. The next section is dedicated to discussing this issue.

4.2.2 Cardinality of the proposition domain

As discussed, Axiom 4 includes an assumption that the proposition domain of Cox's plausibility function should be dense. Let us briefly elaborate on that statement, before discussing its validity.

The CJ approach involves a plausibility function $p : (\mathscr{A} \times \mathscr{A}) \to [0, 1]$ that maps a proposition $(A|B) \in (\mathscr{A} \times \mathscr{A})$ to a real number in $[0, 1]$. Here $\mathscr{A}$ is the set of propositions such as A or B. The density axiom assumes that $\mathscr{A}$ is infinite in size, since only then can $\mathcal{P}_0$ (in the axiom) be a dense subset of $(\mathcal{F}, \mathcal{T})$ as required. Since $\mathscr{A}$ is the set containing the proposition we wish to reason about, we thus require an infinite number of propositions for any problem.

There now arises the matter of whether the plausibility domain should be infinite when dealing with problems that are only relevant to finite sets of propositions. Halpern did not think this reasonable, as he states in his 1996 paper (page 1319):

> *"(...) if we really are interested in a single domain, the motivation for making requirements on the behavior of F on belief values that do not arise is not so clear."*

He repeats this sentiment in a later work (Halpern 1999), where he agrees with Snow's earlier correction on Cox's actual assumptions. In that paper, Halpern explores the possibility of weakening Paris's final axiom (our Axiom 4) while retaining the conclusion that probability theory emerges from the assumptions. He succeeds partially, providing two separate theorems that would each suffice in place of our density axiom, but notes that the differences between either and the density axiom are minute. Importantly, neither solves his concerns regarding the size of the plausibility domain.

It has been argued that density might be a desirable property if one wants a theory that applies to arbitrary domains: a finite domain would necessarily not be sufficient to reason about all possible plausibilities (e.g. when an infinite number of gradations of belief need to be considered). Halpern mentions this motivation in the second of his papers, but dismisses it for excluding finite domains nonetheless. Indeed, in many settings of interest - especially in AI - the plausibility domain is finite (Halpern 1999), so Halpern certainly has a point. Van Horn (2003), however, argues in favour of arbitrary domains on pragmatic grounds (see also footnote 7 of his paper):

> *"[Humanity] would have found it difficult to make any significant progress in mathematics if we had been required to come up with new rules of logic for every new domain we wished to investigate. It is the very fact that we have identified widely-applicable rules of logic, to be used in nearly every domain, that allows us to reason with confidence when entering new conceptual territory."*

Van Horn further argues that finite proposition domains and density need not be mutually exclusive. Considering a finite set of propositions $\mathscr{A}$, one would (correctly) conclude that the number of propositions $(A|B)$ in $\mathscr{A}$ is finite. However, Van Horn argues, the traditional notation $(A|B)$ with A and B propositions hides a dependence on the state of information X. The correct notation to use is then $(A|B, X)$ for any X in the basis of $\mathscr{A}$. Thus, the proposition domain may be finite for a given problem when excluding the state of information, but the plausibility domain of that problem will only be finite if X is restricted to a finite number of states of information.

Snow moreover argues against such a restriction already in his first response to Halpern (Snow 1998). He notes it is often the case that problems to which plausible reasoning is

applied include events that are described by some objective measure. In such a case our assignment of plausibilities should correspond to that measure, which can in principle take any number of values a priori.

Using the concept of a state of information Snow's argument can be elucidated via an example by Van Horn. Consider the proposition A that a particular radioactive isotope will decay in a given time period. Now the assignment of plausibilities $(A|X)$ will heavily depend on the state of information X. As an example, X could be the information about the isotope's half-life, which can be any number of values for an unknown isotope.

Technically, one might argue, an isotope's half-life - and related information - can still only take a finite number of possible values. Let us then consider another argument by Snow (1998) that does not depend on physical measures, as adapted by Van Horn. Consider a right triangle Z of width x and height y, and the proposition B that a point p lies on the vertical leg of Z. Denote by $X(x, y)$ the state of information in which all we know is that p lies on the perimeter of Z. In the case where $y \ll x$ it seems reasonable to conclude that $(B|X(x, y))$ is close to $\mathcal{F}$; the statement that our proposition is false. Vice-versa, $(B|X(x, y))$ should be close to $\mathcal{T}$ for $y \gg x$, and $(B|X(x, y))$ should increase smoothly as y/x varies from zero to infinity (if this were not the case, some triangles Z would have to be excluded from consideration). Thus $X(x, y)$ entails an countably infinite number of different states of information.

Terenin and Draper (2015) remark in the same spirit that:

> *"Jaynes (2003) makes an important distinction between expressions of information that are ontological (e.g., "there is noise in the room") describing the world as it is and those that are epistemological (e.g., "the room is noisy") describing Your information about the world. There is no contradiction in assuming a finite number of world states (ontology), which is certainly true in some problems, and using an uncountably infinite number of propositions to describe Your uncertainty about those world states (epistemology); the latter is a modelling choice that we (and many other Bayesian statisticians) and to be extremely useful."*

A more detailed discussion of the terms epistemology and ontology - plus their potential role in physics - is provided in Chapter 5.

Terenin and Draper, contrary to Snow and Van Horn, mention they agree with Halpern that the restriction of the proposition domain to finite size is *"an unacceptable restriction to the scope of the CJ approach"*. This has motivated them to create a new axiomatisation of the Cox-Jaynes approach that allows for both finite and (un)countably infinite domains. The consequences of this will be discussed in section 4.3.

Finally, as may have been concluded from the above quotation, Jaynes himself was a proponent of finite proposition domains as a rule. He refers to his 'cautious approach' policy regarding infinite sets, considering finite sets the safe harbor on which his desiderata were based and provably consistent (Jaynes 2003, Appendix B.2). An extension to infinite domains, following his policy, may then only be achieved by a well-behaved limit starting from a finite case. That is, one should never attempt to work with infinite sets from the onset. Terenin and Draper consider this approach too cautious, as may be concluded from their seemingly successful attempt at rigorising the Cox-Jaynes approach for (un)countably infinite domains, discussed in the next section.

For now, however, the issue raised in this section seems as of yet unresolved. The restriction of the proposition domain to infinite cardinality is considered unwarranted by

some, while others consider it natural, or even necessary. Under the state of information interpretation by Van Horn, the domain may be considered both finite and infinite simultaneously, depending on whether one includes these information states or not. A possible resolution to the controversy raised here has been provided by Terenin and Draper (2015), whose method simultaneously solves the third point of controversy - countable additivity of the probability function - as well. For that reason, we discuss their solution in the next section.

4.3 Countable additivity of p

With the previous point of controversy seemingly still undecided we turn to Terenin and Draper's attempt at a more rigorous proof of Cox's theorem. Their motivation for the attempt is twofold. First, as discussed, the density axiom by Paris's (essentially equivalent to our Axiom 4) restricts the domain of the CJ approach to infinite sets of propositions. The axiom was necessary to defeat Halpern's counterexample to Cox's approach, but Terenin and Draper considered this domain restriction to be unacceptable. They introduce the notion of 'comparative extendability' to circumvent Halpern's example, while allowing the relevant proposition domain to remain finite if necessary. This is achieved by adding to the finite domain an infinite number of irrelevant propositions, such as the outcomes of fair coin flips. Such an extension is a standard move in works of this type; as pointed out by Halpern (1999), both Savage (1954) and Snow (1998) make a similar argument.

Second, Terenin and Draper wish to extend the CJ approach to include countable additivity of the plausibility function. Previous versions of Cox's theorem, they claim, have never been shown to achieve this. Instead, such approaches only explicitly satisfied finite additivity. The subjective approach to Bayesian probability by de Finetti (1931; 1937) has this same issue, and as such one might consider these Bayesian methods incomplete. In light of Paris's density axiom - which requires infinite proposition domains - such a lack of countable additivity seems especially problematic: one cannot consistently reason about infinite domains using only finite additivity.

Additionally, by what Terenin and Draper call Stone's representation theorem, every Cox-Jaynes Boolean algebra is isomorphic to a Kolmogorov set-theoretic space (Stone 1936). In their words: *"(...) Cox-Jaynes may be regarded as equivalent to Kolmogorov, but in a world in which conditional probability is the primitive concept."* One of Kolmogorov's axioms is the countable additivity of his spaces; thus Stone's theorem gives an indication that the CJ plausibility function should satisfy countable additivity.

This possible connection between the Cox-Jaynes approach and that of Kolmogorov (1933) is quite interesting. The formalism by Kolmogorov is generally accepted among statisticians, and the differences between it and the CJ approach were a reason to doubt the validity of the latter. Stone's theorem closes the gap between these two formalisms, which makes the absence of countable additivity in previous proofs of Cox's theorem even more curious. Terenin and Draper's new axiomatisation may then represent a reconciliation of the Bayesian perspective with Kolmogorov's axioms. Moreover, it is a possible solution to the controversy surrounding the cardinality of the proposition domain - discussed in section 4.2.2 - as it does away with Paris's density axiom.

It is of note that Jaynes had already mentioned that Kolmogorov's axioms and his own formulation of Cox's original theorem need not disagree, although he did not remark on the issue of additivity (Jaynes 2003, Appendix A.1):

> *"(...) we see no substantive conflict between our system of probability and Kolmogorov's as far as it goes; rather, we have sought a deeper conceptual foundation which allows it to be extended to a wider class of applications, required by current problems of science."*

This section is dedicated to a discussion of Terenin and Draper's contribution. Their alternative axiomatisation leading to Cox's theorem will be presented in section 4.3.1, followed by a brief aside on the validity of their claim that previous axiomatisations fail to satisfy countable additivity in section 4.3.2.

4.3.1 A brief reaxiomatisation

In the following we will briefly walk through Terenin and Draper's derivation. Using Stone's representation theorem they rigorise the Cox-Jaynes approach in the language of Kolmogorov's set-theoretic spaces. They construct sequential functions P_1 to P_4 that satisfy an increasing number of rules from probability theory, until P_4 can be recognised as the CJ plausibility function.

Definition TD1: domain. Let Ω be a finite or (countably or uncountably) infinite set, and let $\mathcal{F}$ be a σ-algebra on Ω. A σ-algebra on a set is a collection of subsets on that set that includes the empty subset, is closed under complement, and is closed under union or intersection of countably many subsets.

This freedom in the size of the proposition domain is a first step to solving the controversy of section 4.2.2. It remains to be shown that this new axiomatisation can defeat Halpern's counterexample without resorting to Paris's density axiom, which would again force the domain to be infinite.

Axiom TD1: representation. Let $P_1 : \mathcal{F} \times (\mathcal{F} \setminus \emptyset) \to R \subset \mathbb{R}$ be a function, written using the notation $P_1(A|B)$.

The content of this is essentially our previous Axiom 1, though this already makes explicit mention of some plausibility function P_1. The following axioms require additional properties of P_1.

Axiom TD2: sequential continuity. For nonempty B,

1. Let $A_1 \subseteq A_2 \subseteq \dots$ such that $A_i \nearrow A$. Then $P_1(A_i|B) \nearrow P_1(A|B)$.
2. Let $A_1 \supseteq A_2 \supseteq \dots$ such that $A_i \searrow A$. Then $P_1(A_i|B) \searrow P_1(A|B)$.

In Van Horn's derivation continuity entered as a result of Paris's density axiom (Horn 2003). Here it is assumed, with the justification that as two events become indistinguishable, their probabilities should be the same: the negation of this axiom implies that two events A and A^* exist that can be made arbitrarily close, yet $P(A|B)$ and $P(A^*|B)$ do not converge.

Axiom TD3: product rule. For some function $f : (R \times R) \to R$ and some sets $(U, V, X, Y) \in \mathcal{F}$ involving (A, B, C) it is true that $P_1(AB|C)$ can be written as:

$$P_1(AB|C) = f\left[P_1(U|V), P_1(X|Y)\right] \tag{4.1}$$

Axiom TD4: sum rule. There exists a function $h : R \to R$ such that:

$$P_1(A^C|B) = h\left[P_1(A|B\right] \tag{4.2}$$

These correspond to our Axioms 5 and 3 respectively, and their justifications are similar. Finally, the essential axiom to defeat Halpern's counterexample without resorting to Paris's axiom is one previously mentioned: comparative extendability.

Axiom TD5: comparative extendability. Define the extended probability function

$$P_E : \mathcal{F}_E \times (\mathcal{F}_E \setminus \emptyset) \to S, \tag{4.3}$$

where $S \equiv [\min(R), \max(R)]$ and $\mathcal{F}_E \equiv \mathcal{F} \otimes \mathcal{B}(S)$ is an extension of $\mathcal{F}$ on the extended domain $\Omega_E \equiv \Omega \times S$. Here $\mathcal{B}(S)$ is the Borel σ-algebra and $\otimes$ specifies a σ-algebra on the product space of $\mathcal{F}$ and $\mathcal{B}$. P_E is a function of two pairs of arguments (A, X) and (B, Y) - with $X, Y \in \mathcal{B}(S)$ - that satisfies all the properties of P_1 and the additional property that for $B, Y \neq \emptyset$:

$$P_E\left[(A, X)|(B, Y)\right] = f\left[P_1(A|B), P_U(X|Y)\right]. \tag{4.4}$$

Here $P_U(X|Y) \equiv j\left[\text{Leb}(X \cap Y), \text{Leb}(Y)\right]$ satisfies all the previous axioms. Leb(.) represents the Lebesgue measure on $\mathbb{R}$, and j is some function from $\mathbb{R} \times \mathbb{R}$ to $\mathbb{R}$ normalised such that $P_U(S|S) = P_1(\Omega|\Omega)$ and $P_U(\emptyset|S) = P_1(\emptyset|\Omega)$.

This axiom is rather technical, but the basic idea is that one can always bolt on an unrelated distribution on S describing some hypothetical event to the domain Ω (an infinite number of fair coin flips for example). Such distributions are not relevant for the problem at hand, but they allow the domain to be infinite even when Ω is not. The additional property on P_E guarantees that the original product rule is still satisfied after extra distributions have been bolted on, and this is the essential property that defeats Halpern's counterexample of a function that satisfies all of Cox's axioms but does satisfy the product rule.

Terenin and Draper remark that this axiom is at least as strong as Paris's density axiom from a purely mathematical perspective. They argue, however, that it is much more natural from a philosophical and applied-statistics perspective, which is their motivation for using it. It may be possible that there are more elegant solutions that also rule out Halpern's counterexample, but such a solution has not yet been found (Terenin and Draper 2015, p.13).

From here, most of the derivations are similar to those of Van Horn. Terenin and Draper prove the consistency of P_1 and find a more concrete form of the product rule (our function F). They derive the associativity equation (3.4), considering both the case that the range of P_1 is dense in $R \subset \mathbb{R}$ and that it is not. The first is equivalent to Van Horn's treatment, as Paris's axiom guarantees density. The second case is proven by extending P_1 to P_E, where the latter is dense in R because one can always add an arbitrary number of events from some distribution to the relevant proposition.

The paper then introduces P_2 as the equivalent to our w (equation (3.15)) and then normalises it to form P_3. Finally, a functional equation similar to that for S (equation (3.33)) is derived, and P_4 is identified as some power of P_3, in the same way that Van Horn identified p as some power of w (proposition 9). As such P_4 has been proven to be equivalent to the Cox-Jaynes plausibility function, satisfying both the product rule and the sum rule.

It is clear from the sum rule (more generally, equation (3.34)) that P_4 satisfies finite additivity, which may be denoted

$$P_4\left(\bigcup_{i=1}^{n} A_i|B\right) = \sum_{i=1}^{n} P_4(A_i|B), \tag{4.5}$$

for any positive integer n, disjoint $A_i \in \mathcal{F}$ and nonempty $B \in \mathcal{F}$. However, Terenin and Draper are able to proof countable additivity of P_4 from this statement of finite additivity by combining it with the sequential continuity of Axiom TD2. In essence they prove that finite additivity plus continuity necessarily leads to countable additivity, which is another reason to require countable additivity of the CJ plausibility function. We provide their proof here.

Theorem TD1: countable additivity. For disjoint $A_i \in \mathcal{F}$, P_4 satisfies

$$P_4\left(\bigcup_{i=1}^{\infty} A_i|B\right) = \sum_{i=1}^{\infty} P_4(A_i|B), \tag{4.6}$$

for any nonempty $B \in \mathcal{F}$.

Proof: Let $(A_1, A_2, \dots)$ be a collection of disjoint sets in $\mathcal{F}$ and let B be any nonempty set in $\mathcal{F}$. Define $A_n^* \equiv \bigcup_{i=1}^{n} A_i \subseteq \Omega$ and $A \equiv \bigcup_{i=1}^{\infty} A_i \subseteq \Omega$. It is clear that $A_n^* \subseteq A_{n+1}^*$ for all positive integers n with $A_n^* \nearrow A$, so:

$$P_4\left(\bigcup_{i=1}^{\infty} A_i|B\right) = P_4\left(\lim_{n\to\infty} A_n^*|B\right) = P_4(A|B). \tag{4.7}$$

By finite additivity of P_4 and sequential continuity it is also true that:

$$\begin{aligned}\sum_{i=1}^{\infty} P_4(A_i|B) &= \lim_{n\to\infty} P_4\left(\bigcup_{i=1}^{n} A_i|B\right) \\ &= \lim_{n\to\infty} P_4(A_n^*|B) \\ &= P_4(A|B).\end{aligned} \tag{4.8}$$

Thus, we obtain $P_4\left(\bigcup_{i=1}^{\infty} A_i|B\right) = P_4(A|B) = \sum_{i=1}^{\infty} P_4(A_i|B)$ as required. □

In this way, the work by Terenin and Draper has solved the controversy of the proposition domain, by pointing out the additional problem of countable additivity. Stone's representation theorem proved to be an essential ingredient in indicating this issue, and it is now clear that the CJ approach and Kolmogorov axioms may be considered two sides of the same coin.

4.3.2 Countable additivity in previous proofs

Terenin and Draper's work provides a strong argument for the unification of the CJ approach with Kolmogorov's axioms. However, as their approach was based partly on the wish to construct a countably additive plausibility function, we should note here that it is unclear that previous proofs of Cox's theorem explicitly violated this countable additivity. Indeed, the proof above only required that the plausibility function be finitely additive and continuous in sequence; two properties that Van Horn's axioms also seem to imply. Indeed, Terenin and Draper summarise this aspect of their issue with the current state of the CJ approach as (2015, p.10):

> *"Many authors have tried to circumvent Paris's Axiom, but none of their approaches has been demonstrated to achieve Countable Additivity."*

which is not a claim that the previous approaches violate countable additivity. This aspect of their work may thus not be strictly novel, although the simultaneous circumvention of the density axiom and inclusion of countable additivity likely is.

It seems thus that Terenin and Draper have provided an original axiomatisation that replaces Paris's density axiom with the natural assumptions of sequential continuity and comparative extendability, have likely solved the issue of proposition domain cardinality, and have possibly reconciled the CJ approach with Kolmogorov's axioms.

This concludes the discussion on controversies surrounding Cox's theorem. While the final two points of controversy (cardinality of the proposion domain and countable additivity of p) seem to have been resolved by the work of Terenin and Draper, the first issue (representation of plausibilities by real numbers) remains a valid point of disagreement. Still, while the limitations of Axiom 1 should be recognised - most notably its problems with representing ignorance - it does not seem unreasonable to accept it provisionally. At the very least it may be concluded that Bayesian probability theory is the best one-dimensional candidate for a system of plausible reasoning.

Chapter 5

The mind projection fallacy

It is clear from the previous that the Bayesian view of probability as representing a measure of reasonable belief has strong justifications both in theorems such as Cox's, as well as in practical applications discussed in Chapter 2. An expansion on this idea - called 'probability theory as logic' - was advocated by Jaynes, one of the strongest supporters of Bayesian probability. The intention is to extend Bayesian probability theory to apply to every case of logical inference, so that one may make consistent logical statements even in situations where information is missing. In this sense, Bayesian probability theory may be viewed as a generalisation of deductive logic to cases with incomplete information.

As mentioned previously in Chapter 2, probabilities are not necessarily associated with frequencies in the Bayesian view, but rather represent some (subjective or objective) measure of belief. Jaynes, an objective Bayesian, was of the opinion that probabilities must represent some measure of reasonable belief, and are thus never indicative of actual stochastic processes in nature. He claimed that the belief that probabilities are actually physical is a form of the mind projection fallacy, a is a cognitive bias introduced by him during the Ninth Annual Workshop on Maximum Entropy and Bayesian Methods in 1989 (Jaynes 1990). According to Jaynes, the fallacy occurs in two forms:

1. One's own belief is seen as a real property of nature.
2. One's own ignorance is seen as a property of nature.

The classical example is ancient tribes of humans creating gods to explain natural phenomena. They imagine a weather god, and then ascribe the natural phenomenon of rain to that god's actions. Another example, which forms Jaynes's main motivation for writing the paper, is that of considering probabilities - which merely describe states of knowledge from the Bayesian perspective - to be real properties of nature.

A clear example in physics is that of Bose and Fermi statistics. In undergraduate courses the argument is sometimes made that *"You and I cannot distinguish between the particles: therefore, the particles behave differently than if we could."* (Jaynes 1990). Of course, it would be quite strange indeed if an experimenter's knowledge alone were to affect a given experiment. The statistics that describe identical particles are a feature of one's ignorance: if one knew the positions and momenta of every particle, one would in principle be able to calculate the evolution of the system, no statistics required. That is, classically.

This last caveat remarks on a crucial point for Jaynes's philosophy. Only 'if one knew the positions and momenta of every particle' could the system's evolution be calculated, yet Heisenberg's uncertainty principle forbids this. Quantum mechanics then, seems to pose a problem for Jaynes's statement and the pure Bayesian view that probabilities can

only represent one's state of knowledge. This concern is exacerbated by the probabilistic nature of quantum measurements. If this issue is not resolved, then it seems the Bayesian view can not be a universal interpretation of the nature of probability. Rather, it would seem to only be justified in the limiting cases where classical physics provides an accurate description of the system under consideration. Of course, most applications of statistics exist in this realm, and thus the concern would be mainly philosophical. Nevertheless, the upshot of a deeper discussion will be that the issue is more nuanced than the previous suggests, and may even yield practical applications.

This chapter will thus be dedicated to an examination of Jaynes's specific Bayesian view of probability, especially in the context of physics. As is also demonstrated in Appendix A.2, the Bayesian view has much to offer classical physics. To further show this, and moreover to show the prevalence of some forms of the mind projection fallacy in physics, a few examples of such applications will be provided in sections 5.2 and 5.3.

Another goal of these preliminary discussions is perhaps to show that Jaynes, despite his controversial ideas, is - to be blunt - no crackpot. Indeed, he has made important contributions to both physics and statistics, such as proposing the Jaynes-Cummings model of quantum optics (Jaynes and Cummings 1963), and solving the Bertrand paradox of probability theory (Jaynes 1973). The reason we wish to express this sentiment is that ideas of quantum realism are often discarded out of hand. Certainly it is true that many are not worthy of much discussion, yet others may have some merit and might as such be considered more seriously. Examples of the latter include the many-world interpretation (Everett 1956; Everett 1957), De Broglie-Bohm theory (Bohm 1952), and perhaps some future extension of Jaynes's view. There does not yet exist a scientific consensus on interpretations of the quantum mechanical formalism, especially on the issue of wavefunction collapse, and progress in this aspect of physics is slow. It may be the opinion of some physicists that physics should only be concerned with predictions; i.e. not with questions of the form 'what is actually going on when ... ?'. These people may be correct in thinking this, but it is not clear that the process of answering philosophical questions can not give rise to new tools for prediction. To quote Jaynes himself (Jaynes 1996b):

> *"If we could either succeed in [our attempt to reinterpret quantum probabilities], or prove that it is impossible, we would know far more about the basis of our present theory and about future possibilities than we do today."*

It is therefore the opinion of the author that Jaynes's viewpoint is worth discussing in full. Thus, after presenting a few classical issues that benefit from Jaynes's viewpoint in sections 5.2 and 5.3, we shall discuss his arguments for quantum mechanics as a theory of inference in section 5.4. First however, some terminology that will be crucial to understanding Jaynes's viewpoint is introduced in section 5.1.

5.1 Epistemology and ontology

In Chapter 4 the following commentary on Jaynes by Terenin and Draper (2015) was already brought to attention:

> *"Jaynes (2003) makes an important distinction between expressions of information that are ontological (e.g., "there is noise in the room") describing the world as it is and those that are epistemological (e.g., "the room is noisy") describing Your information about the world."*

This distinction between ontology and epistemology is crucial for understanding the mind projection fallacy. Indeed, one might say that the mind projection fallacy is nothing more than not making this distinction. Note however, that instead of the expressions 'epistemological' and 'ontological' we will elect to adopt the phrases 'epistemic' and 'ontic'. The difference is subtle but relevant, as the former terms respectively relate to the study of various aspects of knowledge and being, where the latter respectively relate to knowledge and being themselves, which is the use relevant for our purposes (Wiktionary 2016).

This section is dedicated to defining these terms and demonstrating their role in the Bayesian view of probability. In section 5.1.1 we will attempt to define the precise content of the terms epistemology and ontology for our purposes. Following that, section 5.1.2 shall treat the first example of the mind projection fallacy in probabilistic reasoning: randomisation.

5.1.1 Models and reality

Let us first explain further what is meant by the terms epistemology and ontology. Taken literally from their Greek roots, the words mean respectively 'study of knowledge' (Dictionary 1996a) and 'study of being' (Dictionary 1996b). These definitions coincide with their use in the above passage: the statement that the room is noisy is one about one's knowledge about the room, while the statement that there is noise is the room is one about the state of the room. Note that there would still be noise in the room if no one were there to observe it, though it could then hardly be said that such a room were noisy, as whether or not a given volume of noise is noisy depends on personal judgment.

Returning to the example of ancient humans creating gods to explain nature, we see that 'the Rain God made it rain' is an ontic statement, though the acceptance of that statement is fully based on one's belief, and thus epistemic. Assigning one's belief about reality that 'the Rain God made it rain' the status of 'this is how reality works' is confusing epistemic and ontic statements: the mind projection fallacy in action.

One could contend that this distinction might also apply to the scientific practice of formulating hypotheses about reality. Jaynes denies this, writing (Jaynes 1990):

> *"(...) the scientist clearly recognizes the creations of his imagination as tentative working hypotheses to be tested by observation; and he is prepared to test and reject a hundred different hypotheses in order to find the right one."*

The scientist then only assigns even epistemic status to a hypothesis once it has been tested beyond all reasonable doubt: models in physics such as thermodynamics, statistical physics, special relativity and Newtonian gravity are only accepted because they measurably describe aspects of reality, in the sense that they predict the outcomes of certain experiments. Such theories must have basis in actual ontic properties of reality, otherwise they could not provide accurate predictions, but they are not themselves complete descriptions of the way nature works fundamentally. Indeed, many classical theories model particles as little balls bouncing of each other, which - while a useful approximation - is not a realistic description. The purpose of physics is essentially to create and test increasingly accurate models, until eventually the unique realistic - ontic - description of reality may be found, if such a finding is at all possible. In the meanwhile however, equivocating models of reality with reality itself is a pure example of the mind projection fallacy.

A common aphorism in statistics - generally attributed to statistician George E.P. Box (Box and Draper 1987, p.424) - is that *"all models are wrong, but some are useful"*. This

describes the mind projection fallacy very well, and holds for any physical models that do not explain literally everything.

Jaynes argues that while the above is well understood by any scientist, many succumb to the mind projection fallacy when probabilities are involved. We elaborate on this claim in the next section.

5.1.2 Probabilistic reasoning and randomisation

Let us illustrate Jaynes's claim that the mind projection fallacy is especially prevalent in the context of probabilities with the example of a coin flip, already explored in section 2.1.

When flipping a (fair) coin, one may say the probability of such a coin coming up heads is a half. This probability is not a property of the coin itself, but rather of one's knowledge of the system. One way too see this, is that this flip could be modelled by classical mechanics if all the relevant forces acting on the coin were known. The outcome of the coin flip could then be calculated, requiring no probabilities. It is only the lack of knowledge of relevant forces that leads one to suppose the flip is random, and should be described by probabilities. This is a useful modelling choice (epistemic), but it is not a statement that reality - classically - runs on probabilities (ontic).

Just as probability, Jaynes argues, randomness is also not a property of reality (Jaynes 2003, section 3.8.1); it is again simply a lack of available knowledge. To illustrate, consider drawing from an urn containing nine white balls, after tossing in one red ball. The probability of drawing the red ball surely depends on a multitude of factors: the size and shape of the urn, the exact way one tossed the red ball in, the elastic properties of the balls and urn, the way in which one reaches into the urn, etc. While it is in principle possible to calculate how one should reach into the urn to grab the red ball, it is wholly unpractical.

Given that this calculation will not be performed, one might consider the probabilistic treatment that the question already hinted at. However, it seems the naive answer of $\frac{1}{10}$ may not be appropriate: usually such probabilities arise from symmetry considerations, in this case that every ball is equally likely to be drawn. However, since the red ball was thrown in last it is likely that is lies on top of the pile, making it easier to grab. One might compensate by altering one's grabbing strategy, but there is no way to know how much compensation is required, besides by doing the calculation.

The solution, Jaynes laconically remarks, is shaking the urn. The shaking has made the problem orders of magnitude harder to solve by calculation. Therefore one now asserts that shaking the urn has made all details irrelevant, so that the problem reverts back to the simple case where the probability is $\frac{1}{10}$. It now remains to invent the term 'randomisation' as a euphemism for 'deliberately throwing away relevant information when it becomes too complicated to handle'.

Jaynes claims this laconic description is necessary as an antidote to some writers on probability theory, who attach a kind of mystical significance to the procedure of randomisation. He admits that the procedure often leads to a useful approximation of the correct solution - and indeed uses it in many of the derivations that follow - but objects to the notion that randomisation somehow makes subsequent equations exact. Randomisation, according to Jaynes, is simply a useful modelling choice, not a physical property. Not appreciating this difference then provides the first example of the mind projection fallacy in the context of probabilistic reasoning.

Indeed, in practice 'random draws' are only approximately random, in the sense that the experimenter who performs these draws lacks the information to manipulate the outcome in a certain way. Having access to such information would defeat the randomness of the experiment, which implies that randomness is indeed a function of the experimenter's knowledge. It is only the absence of relevant information that allows an assignment of the symmetrical probability $\frac{1}{10}$ - as an approximation - in the above experiment.

This has been an introductory note on the distinction between epistemology and ontology in the context of physical models. It has served to supply some background for further discussion of the mind projection fallacy. The following section builds on this background by providing two more mathematical settings that demonstrate that it is - in fact - a fallacy to consider classical probabilities to be ontic elements of nature.

5.2 Inference and causation

The purpose of this section is to show that considering classical probabilities to be ontic is a fallacy. To do this, two situations are described that lead to causality paradoxes when classical probabilities are thought to express real physical causation, rather than inferences. The simplest example of such a paradox appears in the thought experiment of drawing coloured balls from an urn without replacement and comparing forward and backward (in time) inferences (section 5.2.1). A less trivial example following the same logic involves a derivation of the Poisson distribution by Bayesian inference, which we explore in section 5.2.2. These examples are especially illuminating for our later discussion in section 5.4.2.

5.2.1 Bernoulli's urn

The example treated here was published by Jaynes (1989), and is a simple way to demonstrate the mind projection fallacy. Consider an urn containing a number of identical balls. Define the following two propositions:

$I \equiv$ "The urn contains N balls, identical in every respect except that M are red, while the remaining $N - M$ are white. We have no information about the location of particular balls in the urn. The balls are to be drawn blindfolded and without replacement."

$R_i \equiv$ "Red on the i'th draw, $i = 1, 2, \ldots$."

For the first draw we have

$$P(R_1|I) = M/N \tag{5.1}$$

If red was drawn on the first draw, then that changes the contents of the urn to $M - 1$ red and $N - 1$ total. Then for the second draw

$$P(R_2|R_1, I) = (M - 1)/(N - 1), \tag{5.2}$$

such that knowledge or R_1 influences the probability of R_2. Naturally, drawing a red ball first has causal influence on the second draw, so in this case physical causality is correctly modelled by the forward inference. This is called forward inference since it infers the probability of the later R_2 from knowledge of the earlier R_1.

However, consider now the situation where one only knows that the second draw was red and has to give a probability that the first draw was also red. The second draw cannot have a physical influence on the first draw, as the second draw had not been made when the first ball was drawn. Thus, if one thinks probabilities describe physical causality then it must be true that

$$P(R_1|R_2, I) = P(R_1|I). \tag{5.3}$$

This condition however, does not generally hold. To see this, define the proposition

$W_j \equiv$ "White on the j'th draw, $j = 1, 2, \ldots$",

and consider the probability $P(R_2|I)$ of the second draw being red.

$$\begin{aligned} P(R_2|I) &= P(R_2|W_1, I)P(W_1|I) + P(R_2|R_1, I)P(R_1|I) \\ &= \frac{M}{N-1}\frac{N-M}{N} + \frac{M-1}{N-1}\frac{M}{N} \\ &= \frac{M(N-1)}{N(N-1)} = \frac{M}{N} \end{aligned} \tag{5.4}$$

Comparing this to equation (5.1), we see that $P(R_1|I)$ and $P(R_2|I)$ are equal. Note furthermore that the probability $P(R_1R_2|I)$ can be written as

$$P(R_1R_2|I) = P(R_1|R_2, I)\, P(R_2|I), \tag{5.5}$$

and also as

$$P(R_1R_2|I) = P(R_2|R_1, I)\, P(R_1|I). \tag{5.6}$$

It must then follow that

$$P(R_1|R_2, I) = P(R_2|R_1, I), \tag{5.7}$$

which contradicts equation (5.3). It must be concluded that (5.3) is incorrect, since equation (5.7) was arrived at by exclusively using the rules of probability theory. The left-hand side of (5.7) cannot express physical causation, because R_1 happens before R_2. Nevertheless, not only does the outcome of a later event matter to the probability of an earlier event, but the probability is even symmetric in time; $P(R_1|R_2, I)$ and $P(R_2|R_1, I)$ are necessarily equal, and as such it should not matter whether one infers forwards or backwards in time.

From the viewpoint that probabilities represent actual random physical causation in nature this result seems to violate causality. There is however no issue with accepting that information about the second draw can have precisely the same effect on probabilities as information about the first draw, if one takes the perspective of probabilities as representing one's state of knowledge. This is most obvious in the case where $M = 1$, since knowing that the first draw was red then means the second draw must be white, while knowing that the second draw was red must mean that the first draw was white. From the viewpoint that probabilities represent only logical inferences, it is no problem that information from a later time may influence probabilities at an earlier time.

Jaynes wished to construct a system of logic that could deal with uncertainty using probabilities, and is quick to point out that this phenomenon was already present in pure deductive logic. Indeed, by contraposition it is true that if 'A implies B', then '$\neg B$ implies $\neg A$'. Interpreting the former as 'A is the physical cause of B', then, means one should also accept the proposition that '$\neg B$ is the physical cause of $\neg A$', which is not true in general. The upshot here is that logical dependence cannot be understood as causal dependence in logic, and that this is true for relations using probabilities as well.

One might counter that the probability $P(R_2|R_1, I)$ does seem to imply physical causation: indeed, doesn't drawing a red ball first affect the probability of drawing the second red ball? Our response is to refer back to the discussion in section 5.1.2, where the process of randomisation was discussed as an approximation of ignorance. The second probability is affected only to the extent that the experimenter cannot manipulate the outcomes of their draws due to a lack of information. One might for example consider this experiment under the condition that the experimenter knows the position of every ball. Clearly then, the probability of drawing a red ball will be 1 if this is the experimenter's goal regardless of the remaining contents of the urn (assuming there are any red balls left); compare this to the coin flipping robot from section 2.1.2.

There is also the question of what ontic randomness implies of nature: if $P(R, I) = 0.5$ is statement of ontic probability, then the implication is that nature decides whether the experimenter will draw a red or white ball in a truly random manner. That is, the draw is purely random even if all the relevant knowledge is available. Of course this is an absurd claim outside of perhaps quantum mechanics, which we will discuss in section 5.4.

Indeed, the claim made by proponents of non-epistemic probabilities is often not the above, but rather that probabilities seem to contain some ontic information, in the sense that a probability's subject and conditional are related in a physical way. An example is the above $P(R_2|R_1, I)$, where 'red on the first draw' must have a different effect on the second draw than 'white of the first draw' does. That is, given a known number of red and white balls one expects $P(R_2|R_1, I)$ to be smaller than $P(R_2|W_1, I)$. Indeed, the point is not that probabilities cannot be related to physical causality at all, but rather than their values are not indicative of any ontic uncertainty. With sufficient (doctored) information the values of probabilities can vary in whichever way, and it thus appears that there is no room for actual ontic randomness; probabilities are simply a useful modelling choice, and hence epistemic.

This example has been illustrative for showing that probabilities necessarily involve epistemic information, but may not adequately reveal how the mind projection fallacy arises when dealing with more physical models. The next section aims to provide such an illustration in the more apt setting of constructing a Poisson model.

5.2.2 The Poisson distribution

The Poisson distribution is a commonly used to describe the number of times a specific event occurs in a fixed interval of time, if the average rate of such events is known, and each event happens independently of the last. For instance, it has uses in describing radioactive decay when an isotope's half-life is known. Since the Poisson distribution assumes independence of events, one may come to the conclusion that if one's experiment fits a Poisson distribution this indicates that the events are actually independent in nature. Furthermore, one may conclude that the probabilities describe some purely physical dependence between the average rate of events and the events themselves, instead of simply a logical dependence based on the available information. These conclusions however,

are a consequence of the mind projection fallacy according to Jaynes. Let us explore his argument.

Jaynes (1990) argument starts by showing that it is possible to derive the Poisson distribution by inference, directly from the statement that different time intervals should be independent. Accordingly, he defines the prior information as:

$I \equiv$ "There is a positive real number λ such that, given λ, the probability that an event A, or count, will occur in the time interval $(t, t+dt)$ is $p(A|\lambda, I) = \lambda dt$. Furthermore, knowledge of λ makes any information Q about the occurrence or non-occurrence of the event in any other time interval irrelevant to this probability: $p(A|\lambda, Q, I) = p(A|\lambda, I)$."

Now denote by $h(t)$ the probability that there is no count in the interval $(0, t)$, and write the proposition

$R \equiv$ "No count in $(0, t+dt)$."

as the product of two propositions

$R =$ ["No count in $(0, t)$."] $\cdot$ ["No count in $(t, t+dt)$."].

This last equation follows from the independence of different time intervals. Note that the second term is equal to $1 - \lambda dt$, so that we can write:

$$\begin{aligned} h(t+dt) &= h(t)\,[1-\lambda dt] \\ h(t+dt) - h(t) &= -\lambda h(t) dt \\ \frac{h(t+dt)-h(t)}{dt} &= -\lambda h(t) \\ \frac{\partial h}{\partial t} + \lambda h(t) &= 0, \end{aligned} \tag{5.8}$$

where in the last line the limit as $dt \to 0$ has been taken. Using the initial condition $h(0) = 1$ (the proposition 'no counts in the time interval $(0, 0)$' has probability 1) the solution is:

$$h(t) = e^{-\lambda t}. \tag{5.9}$$

Consider now the proposition:

$B \equiv$ "In the interval $(0, t)$ there are exactly n counts, which happen at the times $(t_1, \ldots, t_n)$ with tolerances $(dt_1, \ldots, dt_n)$, where $(0 < t_1 < \cdots < t_n < t)$."

By the independence of time intervals this is a conjunction of $(2n+1)$ propositions:

$B =$ [no count in $(0, t_1)$] $\cdot$ (count in dt_1) $\cdot$ [no count in (t_1, t_2)] $\cdot$ (count in dt_2) $\ldots$ [no count in (t_{n-1}, t_n)] $\cdot$ (count in dt_n) $\cdot$ [no count in (t_n, t)],

and so the probability of B given λ and the prior information I is given by:

$$P(B|\lambda, I) = [e^{-\lambda t_1}] \cdot (\lambda dt_1) \cdot [e^{-\lambda(t_2 - t_1)}] \ldots [e^{-\lambda(t_n - t_{n_1})}] \cdot (\lambda dt_n) \cdot [e^{-\lambda(t - t_n)}]. \tag{5.10}$$

More explicitly, writing (uniquely) $B = dt_1 \ldots dt_n$ we obtain:

$$P(dt_1 \dots dt_n|\lambda, t, I) = e^{-\lambda t}\lambda^n dt_1 \dots dt_n. \tag{5.11}$$

Equation (5.11) puts us in a position to answer a more general question: given λ, what is the probability that in the interval $(0, t)$ there are exactly n counts, whatever the times? The answer to this question should be the usual Poisson distribution. We should sum (5.11) over all times, subject to the condition that $(0 < t_1 < \cdots < t_n < t)$. Since time is taken to be continuous those sums are represented by integrals:

$$P(n|\lambda, t, I) = \int_0^t dt_n \dots \int_0^{t_3} dt_2 \int_0^{t_2} dt_1 e^{-\lambda t}\lambda^n. \tag{5.12}$$

The integrals are straightforward to evaluate and yield

$$P(n|\lambda, t, I) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}, \tag{5.13}$$

which is indeed the familiar Poisson distribution. Usually this distribution is derived as a low counting rate limit of the binomial distribution; an approach that is no less valid. Note that Jaynes's approach makes no mention of physical causation: the derivation depends only on epistemic knowledge of λ and the independence of time intervals.

Indeed, Jaynes rallies against authors who claim that since λ is the only relevant physical agent, it must always be the sole relevant quantity in probability calculations. Indeed, if the Poisson probabilities represent something ontic in nature, and λ alone contains this ontic information, then λ should at the very least appear in any probability distribution derived from this Poisson distribution.

Similarly, he disagrees that verifying the validity of the Poisson distribution with experiments proves that events were indeed causally independent. His argument for the former is similar to that of our urn. The latter he argues against by saying one could write many different computer programs that generate seemingly random data in entirely deterministic ways. In the context of the Poisson distribution this means that the time of the next event is completely determined by the times of the previous events by some complicated rule, in such a way that the long-run frequencies agree with the Poisson distribution.

Thus, Jaynes argues, the events could be correlated in some complicated way that is nearly impossible to discover for any experimenter that does not already know the rule, and whose only tools are statistics and accumulation of more data. The mind projection fallacy might then lead such an experimenter to conclude no rule exists, even though the assumption of independence of events is in such a case the result of their ignorance of that rule, and thus epistemic.

Here we also catch the first glimpse of Jaynes's view of the Copenhagen interpretation of quantum mechanics: just because a deterministic rule that explains quantum probabilities has not been found, does not mean one does not exist. We shall discuss this viewpoint later, but let us first return to the Poisson distribution. In the following, Jaynes's argument against λ as representing something ontic will be presented.

Let $0 < t_1 < t_2$ and let n_1 and n_2 be the number of counts in the intervals $(0, t_1)$ and $(0, t_2)$. We are interested in the joint probability $P(n_1, n_2|\lambda, t_1, t_2, I)$, which can be written out in the following two ways:

$$\begin{aligned} P(n_1, n_2|\lambda, t_1, t_2, I) &= P(n_1|\lambda, t_1, t_2, I)\, P(n_2|n_1, \lambda, t_1, t_2, I) \\ &= P(n_1|\lambda, t_1, I)\, P(n_2|n_1, \lambda, t_1, t_2, I) \qquad (5.14) \\ P(n_1, n_2|\lambda, t_1, t_2, I) &= P(n_2|\lambda, t_1, t_2, I)\, P(n_1|n_2, \lambda, t_1, t_2, I) \\ &= P(n_2|\lambda, t_2, I)\, P(n_1|n_2, \lambda, t_1, t_2, I), \qquad (5.15) \end{aligned}$$

where the last step in each expression follows from the logical independence of time intervals. $P(n_2|n_1, \lambda, t_1, t_2, I)$ corresponds to the forward inference of finding n_2 when n_1 is known, while $P(n_1|n_2, \lambda, t_1, t_2, I)$ corresponds to the backward inference of n_1 when n_2 is known. As the time intervals are logically independent, the forward inference - the probability of $(n_2 - n_1)$ counts in the interval (t_1, t_2) - is given by the Poisson distribution (5.13)

$$P(n_2|n_1, \lambda, t_1, t_2, I) = e^{-\lambda(t_2-t_1)} \frac{(\lambda(t_2 - t_1))^{n_2-n_1}}{(n_2 - n_1)!}, \qquad (5.16)$$

with $t_1 < t_2$ and $n_1 < n_2$. Equation (5.14) then yields:

$$\begin{aligned} P(n_1, n_2|\lambda, t_1, t_2, I) &= P(n_1|\lambda, t_1, I) P(n_2|n_1, \lambda, t_1, t_2, I) \\ &= \left[e^{-\lambda t_1} \frac{(\lambda t_1)^{n_1}}{n_1!} \right] \cdot \left[e^{-\lambda(t_2-t_1)} \frac{(\lambda(t_2 - t_1))^{n_2-n_1}}{(n_2 - n_1)!} \right] \\ &= e^{-\lambda t_2} \lambda^{n_2} \left(\frac{t_1}{t_2 - t_1} \right)^{n_1} \frac{(t_2 - t_1)^{n_2}}{n_1!\,(n_2 - n_1)!} \\ &= e^{-\lambda t_2} \binom{n_2}{n_1} \frac{\lambda^{n_2}}{n_2!} \left(\frac{t_1}{t_2} \right)^{n_1} \frac{(t_2 - t_1)^{n_2}}{(1 - t_1/t_2)^{n_1}} \\ &= e^{-\lambda t_2} \binom{n_2}{n_1} \frac{\lambda^{n_2}}{n_2!} \left(\frac{t_1}{t_2} \right)^{n_1} t_2^{n_2} \left(1 - \frac{t_1}{t_2} \right)^{n_2-n_1} \\ &= \left[e^{-\lambda t_2} \frac{(\lambda t_2)^{n_2}}{n_2!} \right] \cdot \left[\binom{n_2}{n_1} \left(\frac{t_1}{t_2} \right)^{n_1} \left(1 - \frac{t_1}{t_2} \right)^{n_2-n_1} \right]. \qquad (5.17) \end{aligned}$$

The first term is recognised as $P(n_2|\lambda, t_2, I)$, so combining equations (5.15) and (5.17) yields the following for the backwards inference $P(n_1|n_2, \lambda, t_1, t_2, I)$:

$$P(n_1|n_2, \lambda, t_1, t_2, I) = \binom{n_2}{n_1} \left(\frac{t_1}{t_2} \right)^{n_1} \left(1 - \frac{t_1}{t_2} \right)^{n_2-n_1}. \qquad (5.18)$$

In this we recognise the binomial distribution for n_1 successes in n_2 trials, with probability of success t_1/t_2. This is very different from the forward inference $P(n_2|n_1, \lambda, t_1, t_2, I)$ in equation (5.16). In particular, the backwards inference $P(n_1|n_2, \lambda, t_1, t_2, I)$ is independent of λ.

Certainly, Jaynes argues, this result is untenable to those who consider probabilities to describe only random physical dependence; since λ is the only physical causative agent, it must be the only relevant quantity in this view. Yet from the perspective that probabilities only model logical inferences, this is no problem at all.

To reiterate: when reasoning forward to n_2 given n_1, knowledge of λ makes n_1 irrelevant, as in equation (5.16) (specifically, only the difference $(n_2 - n_1)$ matters; i.e. the number

of counts in (t_1, t_2) regardless of the value of n_1). However, in the backwards inference of equation (5.18) knowledge of λ does not make n_2 irrelevant for determining n_1. The opposite is in fact true; knowledge of n_2 makes knowledge of λ irrelevant. Indeed, Jaynes asks, if an experimenter already knows the actual number of events n_2 in an interval, how would they use their knowledge of λ to improve their estimation of n_1 beyond what equation (5.18) provides?

Knowing λ does not fully determine either n_1 or n_2, but only provides probabilities for various possible values; it allows the experimenter to estimate based on the incomplete knowledge in λ. The relevant information contained in knowing the exact value of n_2 (over an interval $(0, t_2)$ that includes t_1) for determining n_1 however, is greater than that contained in knowing λ. In other words, possessing actual data in n_2 makes the previous sampling distributions - that are only dependent on λ - irrelevant; a conclusion that is only sensible when probability distributions depend on the available information.

The derived probabilities must thus necessarily represent logical inferences, claims Jaynes. This is why information about n_2 matters for determining n_1, even though there could be no physical causal relation from n_2 to n_1. In fact, this is reminiscent of the example of last section, where the second draw from an urn could not influence the first draw physically, yet knowledge of the second draw's outcome still affected the probability distribution of the first draw. Indeed, if one were to know what determines λ then λ itself would no longer be necessary to characterise the distribution that describes the process of interest. Instead, this information may lead to a new distribution on the basis of more knowledge, or perhaps a deterministic resolution on the basis of all relevant knowledge.

Note that it is not being denied that probabilities are affected by physical causes, but merely that probabilities are not a fundamental description of reality, at least classically. Accumulation of new knowledge changes probability distributions, and access to all relevant knowledge would render them unnecessary. Physical causes influence probability distributions because they are relevant for a system's description, in the same way that the probability distribution of a flipped coin is affected by the coin's shape. Physical causes shape the way in which probabilities are affected by new information, but the claim is that - fundamentally - those probabilities are a description of that information, not of any truly random process in nature.

In the previous section, the argument was made that the urn probabilities must be fully epistemic, as they vanish when all relevant information is available to the experimenter. A similar argument may work here, but it is dependant on the precise system that is being modelled. One could imagine a classical system that follows a Poisson distribution, such as the number of customers in a store on a given day, and reasonably conclude that such a system would be solvable in principle with sufficient information, such that there is no need for stochastics.

However, a system that requires a quantum mechanical treatment - such as radioactive decay - may not enjoy this measure of determinism. Indeed, quantum mechanics may pose a large obstacle to Jaynes' view that probabilities are never ontic; a discussion of this issue is provided by section 5.4. Classically, at least, it seems that there is a strong argument for viewing probabilities as purely epistemic.

Let us now return to discussing the Poisson distribution and the relations that were derived from it. In the above λ was considered known, which is not always the case in experiment. Jaynes therefore considered the case where λ is not known as well (though he assumed λ to be time-independent for simplicity) (Jaynes 1990). The effect of not knowing λ is that extra integrations over λ are required in many of the above equations. In particular, the

forwards and backwards inferences become:

$$P(n_2|n_1,t_1,t_2,I) = \int P(n_2|\lambda,n_1,t_1,t_2,I)\,P(\lambda|I)d\lambda \tag{5.19}$$

$$P(n_1|n_2,t_1,t_2,I) = \binom{n_2}{n_1}\left(\frac{t_1}{t_2}\right)^{n_1}\left(1-\frac{t_1}{t_2}\right)^{n_2-n_1}. \tag{5.20}$$

The difference between (5.16) and (5.19) represents the information penalty that must be paid for not knowing λ in the forward inference. In the backwards inference, however, there is no penalty to be paid, as knowledge of the exact value of n_2 supersedes knowledge of λ. Indeed, we had already seen that knowledge of λ is irrelevant in the case where it is known, so it seems proper that λ should not matter when unknown either.

This phenomenon is true more generally: sampling distributions are only relevant for making predictions before the data has been gathered. Accumulating data changes the experimenter's state of knowledge, and therefore changes the sampling distributions from the Bayesian viewpoint; this corresponds to updating the posterior distribution. Jaynes gives the example of calculating the probability that 8 or more successes happen in 10 binomial trials. The binomial sampling distribution might assign a probability of 0.32 to this proposition, but if the first three trials yield no successes, then it is certain that there will not be 8 or more in 10 trials.

The purpose of this section was to provide some explicit examples that show considering classical probabilities to be ontic is a fallacy. It was shown that the information interpretation of probabilities is a sufficient description, and that considering probabilities to represent ontic uncertainty can lead to absurd conclusions. This is true for classical systems at least: quantum mechanically - on the contrary - probabilities are often interpreted as containing true ontic randomness in the sense that the theory is not deterministic even in principle. Discussing randomness in quantum mechanics is no small task however, and hence the entirety of section 5.4 is dedicated to this endeavour. The following section 5.3 still explores a classical perspective, though: the view that statistical mechanics in physics may be a theory of pure inference.

5.3 Statistical mechanics and the mind projection fallacy

If one accepts the arguments for classical probabilities as epistemic, then there are some consequences for interpreting statistical theories in physics. Most notably, statistical mechanics - describing systems of uncertain microscopic, but known macroscopic realisation using probability theory - may be considered a theory of pure inference from some initial constraints. The intuition for this is that the randomness associated with statistical behaviour in a system may simply be an expression of an experimenter's ignorance of the system's microscopics. This section attempts to explore this idea of statistical mechanics as inference.

Naturally, it is Jaynes (1989; 1990) who argues for interpreting statistical mechanics as a process of inference, rather than as a theory of physical random behaviour. His primary inspiration may have come from the maximum entropy formalism - discussed in Appendix A.2 - which has enjoyed some success in deriving known results of statistical physics from minimal assumptions and probability theory. He notes that the proposed interpretation solves a number of long-standing philosophical problems in statistical physics as well.

Section 5.3.1 below notes some of these long-standing problems and their solutions according to Jaynes. It also serves to flesh out the idea of statistical mechanics as inference more fully. The section following this (5.3.2) explores some insights Jaynes's perspective yields for understanding the second law of thermodynamics.

5.3.1 Long-standing problems

The probabilities in statistical mechanics have historically been interpreted as the fraction of time that a given system spends in each state. This is fairly reminiscent of the frequentist interpretation of probability: the probability represents a fraction of time in the long run. Indeed, the standard interpretation of a statistical ensemble as a large number of virtual copies a given system, each in a different state, is very similar to the frequentist approach to probability theory. This interpretation however makes it impossible to reason about systems in transitory states using statistical mechanics (Nathaniel 2012). Furthermore, some theories in this framework rely on ergodic hypotheses (essentially: 'ensemble averages are equal to time averages'), which some consider unwarranted and others consider debunked (Jaynes 1967; Jaynes 1978).

One may then wonder whether a Bayesian approach to statistical physics might be more appropriate. Such a theory would consider the probabilities of statistical mechanics to describe how likely a system is to find itself in a particular state at a given time. Comparable to how the Bayesian perspective in section 2.1.2 allowed for a description of the single coin flip - where the frequentist perspective would not - so does the Bayesian interpretation here allow for probabilistic statements without the necessity of an ensemble. Under this interpretation, any probability distribution - such as the microcanonical ensemble - describing a system refers only to possible states of that single system.

Jaynes further contends that the historic prevalence of the ensemble interpretation may have been the consequence of misunderstanding the work of Gibbs. The idea of an ensemble in statistical physics is already present in works of Boltzmann (1882a; 1882b; 1882c) and Maxwell (1882), though it is often thought that Gibbs invented this notion. Jaynes remarks that Gibbs in fact de-emphasised the importance of the ensemble, quoting him (Gibbs 1902) and noting a hint of cynicism in the following:

> *"It is in fact customary in the discussion of probabilities to describe anything which is imperfectly known as something taken at random from a great number of things which are completely described."*

Furthermore, Gibbs remarks that, if we wish to avoid referencing an ensemble of systems, we may recognise that we are merely talking about *"the probability that the phase of a system falls within certain limits at a certain time (...)"*. Jaynes thus observes that Gibbs recognised the ensemble as an ad hoc device invented for the purpose of thinking of probabilities as frequencies. Constructing such an interpretation is of course possible for any probability, but it seems Gibbs recognised that that statistical mechanics really only describes imperfect knowledge about a single system.

The probabilities themselves are then again a measure of the experimenter's ignorance; a tool to make inferences about systems on the basis of incomplete information that is the observable macrostate. In principle one may gather sufficient information to uniquely determine the microstate that underlies a given observed macrostate, at which point the necessity for a probabilistic description vanishes.

Naturally, Jaynes argues for the ignorance interpretation, and notes that it solves a number of long-standing philosophical problems as well. As an example of such a problem, he

quotes Pool (1989) quoting a worker in statistical mechanics as stating as one of its long-standing problems:

> *"Where does the randomness necessary for statistical behavior come from if the universe is at heart an orderly, deterministic place?"*

One may critique the qualifier 'deterministic' by pointing to the randomness of quantum mechanics, but it can hardly be denied that classical physics has demonstrated a tendency of macroscopic behaviour to be deterministic. Indeed, it may very well be true that the universe is at heart not deterministic (in the sense that wave-functions collapse in a truly random way for example), but that does not mean that established deterministic theories of macroscopic physics give wrong predictions for the domains in which they are (approximately) valid: i.e. Newtonian gravity still correctly predicts that apples fall, even though it is superseded by general relativity, which may be in the future superseded by some unknown theory of everything.

Certainly, the behaviour mentioned in the above quote may seem rather problematic from the perspective that probabilities are indicative of physical uncertainty. This problems disappears, however, when taking the viewpoint that statistical physics is merely a theory of inference from incomplete information. There is no need to explain this statistical behaviour, because the randomness is not physical: it is simply an expression of one's ignorance about the degrees of freedom in the microstates underlying a macrostate. Compare this with the coin flip discussed in section 2.1.3 that has an in principle predictable outcome, but is described by probabilities because the relevant information is not available in practice.

One might then wonder how it is possible that macroscopic physics works so well. To explain this objection, consider a system in a given microstate at some time t_1. At that time the system is observed to be in a certain macrostate M_1. This macrostate contains some information, such as the system's pressure and temperature. The system will evolve in some way until at time t_2 it is in the macrostate M_2. This behaviour is predicted by phenomenological thermodynamics, a macroscopic theory with no special regard for the underlying microscopics: how can such a theory accurately predict the evolution of a system that is dictated by degrees of freedom it does not even consider? Stated another way; there are many different microstates that could underlie M_1 and these all evolve according to some microscopic laws. There is a priori no reason to suppose that these evolved microstates all give rise to the same macrostate M_2. How then can the evolution $M_1 \to M_2$ be predictable by macroscopic laws?

Jaynes addresses this issue already in his first work on maximum entropy (1957) and later more thoroughly in a paper titled 'Macroscopic Prediction' (1996). A summary can be found in the work of Kuić et al. (2012), from the perspective of maximum entropy inference. Essential to the argument is a principle of macroscopic uniformity; the empirical fact that definite predictions of macroscopics are only possible when an overwhelming majority of possible microstates give rise to the same macroscopic behaviour.

This is not a new idea, as Kuić et al. readily note: *"In somewhat different context this property is recognized as the concept of macroscopic determinism, whose precise definition involves some sort of thermodynamic limit."*

Jaynes provides another example of confusion arising from the idea that probabilities in statistical mechanics are ontic, referring to Mark Kac (1956) who *"considered it a major unsolved problem to clarify how probability considerations can be introduced in physics"*. Kac stated that he could not understand how one can justify the use of probability theory the way Boltzmann used it to derive the Boltzmann equation. In his attempts the problem

was either under- or over-determined, and he could only get to Boltzmann-like equations in some idealised model where the n-particle distribution factorised into a product of single-particle distributions (see Jaynes 1990 for more on Kac's approach). He thus wrote:

> *"(...) it is philosophically rather peculiar. Because if you believe in it you must ask yourself why nature prepares for you at time zero such a strange factorized distribution. Because otherwise you can't get the Boltzmann equation."*

Here we see the mind projection fallacy in Kac's reasoning as well. Indeed from his perspective it should seem rather curious that the particle distribution factorises, i.e. that an n-particle system can be described as n single-particle systems with no interaction between them. As one may expect, this is not an issue from Jaynes's viewpoint. He comments that his original reply to Kac did not have the intended result, and proceeds to write down what his current - more detailed - response would be (Jaynes 1990):

> *"The probability distributions in phase space used by Maxwell, Boltzmann, and Gibbs are not realities existing in Nature; they are descriptions of incomplete human information about Nature. They yield the best predictions possible from the information contained in them. Probability distributions are not "right" or "wrong" in a factual sense; rather, some distributions make better predictions than others because they contain more relevant information. With a factorized distribution, getting knowledge of the position of one particle would tell us nothing about the position of any other. But at soon as the particles interact, knowing the position of one does tell us something about where the others are. Therefore the probability distributions which lead to the best physical predictions for interacting particles are not factorized."*

Jaynes further comments that the Boltzmann distribution does in effect suppose factorisation, and that is precisely the reason it cannot fully describe the relevant interactions between particles. Gibbs's distribution yields more accurate predictions than Boltzmann's - because it contains more relevant information about interactions - and indeed this distribution is non-factorisable.

In this section some examples of philosophical problems in statistical physics were provided that seem to arise from the mind projection fallacy. Furthermore, the idea of statistical mechanics as a theory of incomplete information was introduced more thoroughly. The next section uses these ideas to provide a better understanding of the second law of thermodynamics.

5.3.2 The second law of thermodynamics

The previous considerations are not the full extent of the applications for a theory of statistical mechanics as inference. In this section we will discuss its role in understanding the second law of thermodynamics: the phenomenological rule that entropy in a closed system tends to increase.

Consider a Carnot heat engine. The goal of such an engine is to concentrate heat - energy spread in an unknown way over any number of microscopic degrees of freedom of a system - into a single degree of freedom, such as the motion of a piston, and as such do work. The efficiency η of a Carnot engine working between upper temperature T_u and lower temperature T_l is given by

$$\eta = 1 - \frac{T_u}{T_l}, \tag{5.21}$$

and is an upper limit of the efficiency of heat engines. In practice, heat engines are not perfectly reversible and thus their efficiency is further limited by the second law of thermodynamics.

Jaynes (1989) argues that the reason for this limitation on efficiency by the second law is that the engine must work reproducibly; that is, reliably, not just when the microstates of the reservoir and engine happen to be right. Indeed, this is another formulation of macroscopic uniformity already mentioned in section 5.3.1 (Kuić et al. 2012).

No experimenter can make a heat engine that works reproducibly (i.e. whenever they want it to given some macroscopics) without a temperature difference in its reservoirs, since they cannot control the microstate underlying the macroscopics. The only tool available to the experimenter is changing the macroscopics, such as temperature and pressure. Associated with such a macrostate is the entropy, which essentially counts the number of possible microstates underlying a macrostate. According to Jaynes however, entropy is not a property of the system under consideration, but rather of the description of that system. Indeed, the entropy of any system of a known microstate is zero, because the system is fully described by that microstate. It is only the fact that an experimenter generally cannot access the microstate of a system that the description in terms of macrostates and entropy has any use at all.

Particular support for this view is provided by Liouville's theorem, which describes the time evolution of phase space distributions. The microstate underlying a macroscopic configuration M_i could be anywhere in some phase volume W_i that contains all microstates compatible with the description M_i. For the engine to work reliably, there must be some reproducible process by which these microstates evolve to give rise to a final macrostate M_f such that work has been done along the way.

One of the consequences of Liouville's theorem is that the possible final microstates of a reproducible process cannot be concentrated in a phase volume W_f that is smaller than the initial W_i, and it thus describes the limiting factor for macroscopic experiments. It is for this reason, claims Jaynes, that we cannot reproducibly go to a final macrostate M_f that is associated with a phase space volume W_f smaller than W_i. As he puts it (Jaynes 1989):

> *"The inequality $W_i \leq W_f$ is a necessary condition for any macroscopic process $M_i \to M_f$ to be reproducible for all initial microstates in W_i."*

To now elucidate the connection with entropy, note that the entropy of a macrostate M is essentially given by $S = \ln W(M)$, such that the statement $W_i \leq W_f$ is equivalent to $S_i \leq S_f$, which is just the second law. Indeed, the second law is simply the statement that systems tend toward equilibrium, and it is well known that the equilibrium state M has the largest phase space volume $W(M)$, often by orders of magnitude (Appendix A.2) (Jaynes 1996a).

The critical point here is that entropy is not a function of the physical state, but rather a function of a description of that state; it is the information contained in the experimenter's description. Not recognising this distinction is another form of the mind projection fallacy; confusing one's own knowledge with a property of nature.

Essentially then, Liouville's theorem states that for any system that reproducibly moves from a state satisfying some description M_i to a state satisfying some other description M_f it must be the case that the phase space of M_f is larger than that of M_i; in other words, the entropy of M_f must be larger than that of M_i. The second law of thermodynamics is thus not so much a law of physics as one of inference, though of course those inferences

are very much dependent on what happens physically.

Jaynes provides an analogy that further demonstrates the role of Liouville's theorem. One can easily pump water from a tank of volume V_1 into a larger tank of volume $V_2 > V_1$, but not into a smaller tank of volume $V_3 < V_1$. Any particular tagged water molecule can thus be reliably moved to the larger tank, but not to the smaller tank. The latter could however still happen, and has a probability of success somewhere around V_3/V_1. In this analogy the water tanks represent the macrostates, their volumes represent the phase space volumes, the tagged molecule corresponds to a specific microstate, and the incompressibility of water models Liouville's theorem.

Certainly, it is still possible that something non-reproducible happens by chance, just as it is possible that some process decreases entropy, or that the heat engine without temperature difference generates work. However, the number of degrees of freedom that need to conspire for this to happen is so large that it in practice never will in macroscopically large systems. Indeed, Jaynes notes that if the entropy of a description M_f of some macroscopic system is smaller than that of M_i by 1 (microcalorie / room temperature), then the probability of moving from a random state described by M_i to one described by M_f is less than $\exp\left[-10^{15}\right]$ (Jaynes 1984b). In other words, a heat engine operating with a cold sink at room temperature may extract an extra microcalorie of energy more than the second law allows, but the probability of this happening is upper bounded by $\exp\left[-10^{15}\right]$.

A proper characterisation of entropy thus requires one to specify the degrees of freedom one is working with (see also Appendix A.2.1). A situation one might encounter experimentally is that the thermodynamic entropy of the specified system spontaneously decreases due to some external influence. For instance, if an experimenter's system S_1 comprises a reservoir of temperature T_1, and this system comes into contact with another - outside - system S_2 of temperature $T_2 < T_1$, then energy (heat) will start flowing out of S_1. For most systems, lower energy corresponds to fewer possible microstates, and as such the entropy of S_1 will decrease once the experimenter notices that energy has left the system, seemingly in violation of the second law and Liouville's theorem.

The common response is that the second law only applies to a full description of both systems (and any other system also in contact) and is therefore safe from this argument. Indeed, two systems of different temperatures in contact will over time exchange heat and move to a description of higher combined entropy.

One might however wonder whether such an experiment contradicts the claim that phase space volume cannot decrease in a reproducible process: entropy - and hence phase space volume - is by our own definition definable for any system, and the phase space of S_1 has gotten smaller in this process. The key insight is that such a process is not reproducible by the experimenter, precisely because S_2 is not part of the degrees of freedom the experimenter is considering: that is, the experimenter has no influence over S_2 or any other systems outside of S_1, and hence cannot reproduce this process. Indeed, the fact that the experimenter can only reproduce this process if they have access to both S_1 and S_2 is precisely the reason that the second law only works for combined systems.

This concludes the present discussion of statistical mechanics as a theory of incomplete information, and also our classical treatment of the mind projection fallacy. It seems that there are strong arguments in support of classical statistical mechanics as a theory of inference. If nothing else, this view provides an intuitive interpretation of entropy as information, dependent on a particular description of state. Appendix A.2 discusses the information interpretation of entropy from the viewpoint of Shannon information theory, which provides additional insights on statistical mechanics as a theory of inference.

The next section attempts to apply the ideas of this chapter to quantum mechanics; in particular, it explores the claim that the Copenhagen interpretation is another form of the mind projection fallacy, and the possibilities for quantum mechanics as a theory of inference.

5.4 Quantum mechanics and the mind projection fallacy

In the previous applications, probability theory represented the process of reasoning from incomplete information. The predictions made by this process are the best that can be made from the available information, as - by Cox's theorem - probability theory is the unique method of correct plausible reasoning. Appendix A.2 will further show that a specific application of probabilistic reasoning can reproduce known results in statistical physics and this suggests that statistical theories could simply be applications of probability theory as logic.

Jaynes wanted to interpret quantum mechanics (QM) in a similar way. As is well known, knowledge of a pure quantum state ψ does not allow an experimenter to predict all possible experimental results deterministically, reminiscent of (but not equivalent to) how an observable macrostate in statistical mechanics does not contain all information of a system. It might then be possible to reinterpret quantum mechanics as a theory of inference; making the best predictions from the partial information in one's possession when ψ is known (Jaynes 1996b).

Jaynes's argument for the Copenhagen interpretation as an application of the mind projection fallacy will first be discussed in section 5.4.1. This will lead to an examination of Bohr's and Einstein's historic argument that culminated in the construction of the Einstein-Podolsky-Rosen (EPR) thought experiment. Section 5.4.2 will examine this thought experiment and the related theorem by Bell more closely, and note possible options for a local deterministic theory of quantum mechanics. Many of the remaining avenues for such theories will be shown to be closed, but it seems the possibility for quantum mechanics as a theory of inference may persist. A discussion of this avenue for QM as inference is provided in this chapter's closing section 5.4.3.

5.4.1 The Copenhagen interpretation

According to Jaynes, the Copenhagen interpretation of QM is a prime example of the mind projection fallacy in physics (Jaynes 1989). It states that physical systems generally have no definite properties before being measured. From the wavefunction one can calculate a probabilistic description of measurement outcomes, and the act of measurement causes the system to collapse into one of the possible values; a process termed wavefunction collapse. Thus the interpretation claims that the state vector ψ is a complete description of reality in the sense that nothing more can be known in principle. Jaynes contends that this last claim is the mind projection fallacy in its second form: one's inability to fully describe the system is equated with nature being nondeterministic; one's own ignorance is seen as a property of nature; the probabilities of QM are understood as causal relations, instead of as logical relations that describe incomplete information. We note here that Jaynes's claim is not one against the mathematical formalism of QM, rather it only asserts that the wavefunction is not a full description of a system.

The Copenhagen interpretation has been heavily criticised, but remains one of the most commonly taught views. That this is the case may be mainly due to the facts that it has

historical significance, and that no consensus on a 'correct' interpretation exists (Wimmel 1992). Perhaps the most famous debate regarding the Copenhagen interpretation has been that between Bohr and Einstein. Bohr was a proponent of the theory, and held that the state vector ψ is a complete description of reality in the previous sense. Einstein, like Jaynes, thought this completeness claim was an unsupported addition and created thought experiments in an attempt to refute Bohr's claims, the most famous of which is now associated with the EPR paradox (Jaynes 1989). Common knowledge is that Einstein lost this debate when Bell's famous inequalities were experimentally violated. Let us then explore this debate in more detail.

Bohr's and Einstein's positions on the meaning of the state vector were analysed by L. Ballentine (1970), who claimed that Einstein's view is often misrepresented. According to him, Einstein's supported a view he terms 'the statistical interpretation', which holds that a pure state represents an ensemble of similarly prepared systems and is thus an incomplete description of an individual system. Jaynes claims that 'virtually all physicists who do real quantum-mechanical calculations interpret their results in the sense of Einstein' (Jaynes 1989). Indeed, Ballentine states that a major aim of his paper is to show that the interpretation of a pure state as providing complete description of an individual system is unnecessary for quantum mechanics and leads to serious difficulties (Ballentine 1970, section 1.2).

Note here the resemblance with the ensemble interpretation in statistical mechanics, which in the previous sections was expanded to an interpretation of inference from incomplete information. This similarity between the classical statistical theory and the quantum mechanical formalism may provide an indication that a similar approach is possible here.

In the simplest terms then, it was the opinion of Einstein that QM was incomplete and that the missing parts could be found, while Bohr held the position that a more complete description was impossible. Jaynes, however, claims that their difference of opinion is not so fundamental as this characterisation might suggest, and that they were merely thinking on different levels. Einstein's position that quantum mechanics is incomplete is ontic, describing physical reality. He believed there was a more complete description of nature available, and that this description can be found. Bohr's perspective of completeness on the other hand was epistemic, describing one's information about reality. He believed that even if there is some more ontic reality underlying QM, we would not be able to find it and make predictions from it.

Physics has traditionally been concerned with construction increasingly accurate models of ontic aspects of nature, and it is for this reason - Jaynes claims - that physicists have had difficulty reading Bohr. J.C. Polkinghorne (1986) independently came to this same conclusion, quoting Bohr as saying:

> *"There is no quantum world. There is only an abstract quantum physical description. It is wrong to think that the task of physics is to find out how nature is. Physics concerns what we can say about nature."*

The notion of a real physical situation was then simply not present in Bohr's writing. Jaynes characterises Bohr's answers to questions of the form *"What is really happening when ...?"* as evasive, and notes that E.P. Wigner was also aware of this. To wit, Wigner - during a colloquium talk at Washington University - remarked (Wigner 1974):

> *"These Copenhagen people are so clever in their use of language that, even after they have answered your question, you still don't know whether the answer was 'yes' or 'no'!"*

While Jaynes's perspective makes Bohr's writing more understandable, Jaynes does not agree with Bohr that physics has no ultimate ontic goal. He commends Bohr for appreciating the fact that *"any theory about reality can have no consequences testable by us unless it can also describe what humans can see and know"* (Jaynes 1996b), as certainly we are limited by the information we can gather. He is however of the opinion that the ultimate goal of science is to learn about the reality that *"continues to go about its business according to its own laws, independently of what humans think or do (...)"* (Jaynes 1996b), that is; the search for ontic properties of nature.

This may seem surprising considering the emphasis Jaynes places on the mind projection fallacy, but we note that it only warns for confusing epistemic and ontic properties. A so-called Theory of Everything is a model that may also be a complete description of nature, and models that precisely correspond to reality are not forbidden.

Furthermore, Jaynes's position regarding the mind projection fallacy is primarily concerned with the nature of probabilistic statements. He does not agree that QM provides sufficient indication that nature is actually random, and therefore agrees with Bohr that QM probabilities are primarily epistemic. Jaynes does note an important caveat that the QM formalism must also contain some correct ontic elements: *"(..) there has to be something physically real in the eigenvalues and matrix elements of the operators from which we obtain detailed predictions of spectral lines"* (Jaynes 1989). We will elaborate on this view in section 5.4.3.

Bohr furthermore claims that nothing more can be known even in principle - i.e. his claim of completeness - and this Jaynes cannot abide. Indeed, the success of epistemic probabilities in describing the macroscopics of statistical mechanics (see also section A.2 - themselves determined by more fundamental microscopics (e.g. the laws governing particles) - leads Jaynes to suspect its parallel in quantum mechanics: the epistemic probabilities of Bohr's interpretation are indicative of a deeper ontic structure, which Einstein wished to find when he refused to accept Bohr's claims of completeness.

The Bohr-Einstein debate is often thought to have conclusively been resolved by Bell's no-go theorem of 1964. In the next section we continue the debate and attempt to show a gap in Bell's work, which may allow for a different interpretation.

5.4.2 The EPR experiment and Bell's theorem

Let us return to a discussion of Bohr-Einstein debate, which culminated into the formulation of the EPR thought experiment. The conclusion of EPR has often been described as a paradox, though Ballentine assures the reader that it is only paradoxical to the extent that one might not have expected the conclusion. In the briefest sense, the EPR thought experiment is in fact a theorem that states the following two statements are incompatible (Ballentine 1970):

1. The state vector ψ provides a complete and exhaustive description of an individual system.

2. The real physical conditions of spatially separated (non-interacting) objects are independent.

We shall term the first statement 'completeness' and the second 'locality'. The importance of the EPR theorem was that it showed for the first time that acceptance of completeness demands rejection of locality, and vice-versa. This rather surprising conclusion in no way

contradicts the mathematical formalism of QM, as is sometimes thought (Ballentine 1970). Rather, it was only intended to show the difficulties with the Copenhagen interpretation.

Clearly, Einstein favoured locality in the above. Bohr seems to have been a proponent of completeness, though Jaynes would likely contend that Bohr's thinking was on the epistemic level, and thus not necessarily applicable to the statements of the EPR theorem. Indeed, Jaynes notes (1989): *"Bohr had never claimed that [the QM state vector] was [a representation of the real physical situation of a system], although his strange way of expressing himself often led others to think that he was claiming this."*

Nevertheless, the theorem has interesting implications for the QM formalism. Without any foreknowledge of quantum mechanics, it is locality that would likely be considered most reasonable. This would mean that the quantum mechanical formalism is an incomplete description of individual systems, and may rather describe a statistical ensemble of systems.

A natural next step in furthering our understanding of physics would then be to find a more complete theory; i.e. something underlying the statistical representation of QM. Such theories are often called 'hidden variable theories', because they assert the existence of unknown degrees of freedom that deterministically produce the QM probabilities.

This brings us to the famous paper of Bell, in which he addressed the EPR argument and attempted to show that no local hidden variable theory can reproduce all probabilities of QM (Bell 1964). The assumption of locality is essential to Bell's argument, and it is obvious why Bell made it; he was responding to Einstein, Podolsky and Rosen who favoured the locality of QM over its completeness. Indeed, there had already been proposals for non-local hidden variable theories prior to 1964, such as the De Broglie-Bohm pilot wave theory (Bohm 1952), and the many-worlds interpretation (Everett 1956; Everett 1957).

A simple realisation of the EPR experiment provided by Bohm (1957) is one in which two experimenters each have a Stern-Gernlach apparatus used to measure electron spins. The spin-$\frac{1}{2}$ electrons in question, denoted A and B, were in a pure joint singlet state with perfectly anti-correlated spins in the past, and have been separated by some spin-independent interaction before being moved far apart. Thus, the spins of A and B remain perfectly anti-correlated, but are now space-like separated; i.e. causal relations between the two are now forbidden. Following Jaynes (1989) and Bell (1964), $P(A|a)$ denotes the probability that spin A will be found in the direction of the unit vector a after measurement, and similarly we denote by $P(B|b)$ the probability of measuring spin B to be in direction b. Einstein, Podolsky and Rosen then showed that, after measuring A in any direction, the value of B in that same direction can be predicted with certainty.

Since there can be no causal relation between the two, they concluded that the results of the possible measurements on B must have been predetermined by the physical situation at B; the measurement at A should be irrelevant. Since the joint singlet state wavefunction does not allow for deterministic prediction of the measurement outcomes at A and B, EPR reasoned it must be an incomplete description of the physical situation.

Bell in his paper defines a set of local hidden variable theories in which A and B also depend on some hidden variable λ. He notes that it does not matter for this argument whether λ is a single variable or set of variables, a set of functions, or whether it is discrete or continuous. He then writes the following identity for the joint distribution $P(AB|ab)$:

$$P(AB|ab) = \int P(A|a,\lambda)P(B|b,\lambda)P(\lambda)d\lambda. \tag{5.22}$$

Bell shows that such a distribution cannot reproduce the probabilities later measured in Stern-Gernlach and similar experiments (notably Freedman and Clauser 1972; Aspect et al. 1981; Aspect et al. 1982a; Aspect et al. 1982b and most recently Abellán et al. 2015), which then rules out local hidden variable theories. He did this by devising an inequality that functions as a constraint on coincidences in the chosen EPR experiment. This constraint must be satisfied if there exist underlying local hidden variables that determine outcomes of quantum mechanical processes.

Indeed, if one interprets probabilities as ontic statements denoting physical causality, then equation (5.22) seems to accurately represent the locality constraint of EPR. Jaynes contends however that this is a result of the mind projection fallacy. Referring back to the example of the urn from section 5.2.1, he argues that probabilities represent only epistemic inferences. While - in a local theory - A cannot causally depend on B and vice-versa, there is no issue with a logical dependence between these variables. Indeed, proper application of the product rule tells us that:

$$\begin{aligned} P(AB|ab) &= \int P(AB|a,b,\lambda)P(\lambda|a,b)d\lambda \\ &= \int P(A|B,a,b,\lambda)P(B|a,b,\lambda)P(\lambda|a,b)d\lambda \\ &= \int P(A|B,a,b,\lambda)P(B|b,\lambda)P(\lambda)d\lambda \end{aligned} \tag{5.23}$$

Where in the last step we have used the assumption that knowledge of the experimenters's choices of direction a and b grants no information on λ. Additionally, note that $P(B|a,b,\lambda) = P(B|b,\lambda)$ such that knowing a alone does not change the probabilities on the value of B. However, note also that $P(A|B,a,b,\lambda)$ cannot be reduced in the same manner, since knowing both the direction of measurement and the measurement outcome does influence what one knows about A, as AB is anti-correlated.

Jaynes thus notes that the problem with Bell's equation lies in his factorisation of $P(AB|a,b)$. For this reason it is no wonder that Bell could not reproduce QM probabilities, says Jaynes.

Indeed, this same issue with Bell's probabilities was raised by Colbeck and Renner (2011). They write the following, where we have replaced their notation with that used by Bell and Jaynes:

> *"To quote Bell [2], locality is the requirement that "...the result of a measurement on one system [is] unaffected by operations on a distant system with which it has interacted in the past..." Indeed, our non-signalling conditions reflect this requirement and, in our language, the statement that $P(AB|a,b,\lambda)$ is non-signalling is equivalent to a statement that the model is local (see also the discussion in [28]). (We remind the reader that we do not assume the non-signalling conditions, but instead derive them from the free choice assumption.) In spite of the above quote, Bells formal definition of locality is slightly more restrictive than these non-signalling conditions. Bell considers extending the theory using hidden variables, here denoted by the variable λ. He requires $P(AB|a,b,\lambda) = P(A|a,\lambda)P(B|b,\lambda)$ (see e.g. [13]), which corresponds to assuming not only $P(A|a,b,\lambda) = P(A|a,\lambda)$ and $P(B|a,b,\lambda) = P(B|b,\lambda)$ (the non-signalling constraints, also called parameter-independence in this context), but also $P(A|B,a,b,\lambda) = P(A|a,b,\lambda)$ and $P(B|A,a,b,\lambda) = P(B|a,b,\lambda)$ (also called outcome-independence). These additional constraints do not follow from our assumptions and are not used in this work."*

Colbeck and Renner however also note that Bell's methods may be correct if he is only concerned with theories that are deterministic given the hidden variables, since in such cases knowledge of λ and a should be sufficient to determine the value of A, i.e: $P(A|B, a, b, \lambda) = P(A|a, \lambda)$. Presumably Jaynes was searching for a deterministic theory, and it is unclear how he would respond to this criticism.

He does however present a second argument against the idea that Bell has conclusively proven the impossibility of local hidden variable theories. Jaynes noted that while Bell was quite general in his definition of λ he did not include the possibility that it depends on time, leaving open the option of time-dependent local hidden variable theories even if Bell's other assumptions proved correct. Somewhat ironically, this argument is countered in the very same paper by Colbeck and Renner. Interestingly, work by Stephen Gull of which Jaynes became aware at the same meeting where he presented his paper on Bell excluded this option as well.

Gull's work is reintroduced in Richard Gill's paper 'Time, Finite Statistics, and Bell's Fifth Position' (2003). In this a pair of classical computers is tasked with recreating probabilities predicted in a Bell-CHSH delayed choice experiment (Clauser-Horne-Shimony-Holt (1969), after the authors of a Bell-like inequality that is often accredited to Bell himself). It is shown that the classical computers can never produce correlations that will violate the CHSH inequality in the long run, even using time-dependent strategies, thus excluding time-dependent local hidden variable theories.

Gill compares the method to playing roulette: the rules of roulette prevent any classical computer algorithm from devising a strategy that will beat the house in the long run. Gill quantifies exactly what the 'long run' is using the theory of supermartingales, and as such bounds the likelihood of seeing the CHSH inequality violated by chance by a classical algorithm.

A similar analogy is provided by Russell O'Connor (2015) in his 'Bell's Casino game'. This game provides an example of a gamble that in the long run cannot be won classically, but has positive expected value when the players are allowed to make use of entangled qubits. This example too shows that quantum mechanical measurements are qualitatively different from classical ones.

Colbeck and Renner go one step further, and attempt to show that no quantum hidden variable theory can produce more informative predictions than current quantum mechanics. Their result may be interpreted as a strengthening of Bell's original theorem to include the non-local case. Thus, while they agree with Jaynes on some of his criticism of Bell, it seems that Jaynes's position that there must be some local hidden variable theory underlying QM is untenable.

There remains one explicit loophole in all the above results, namely that they depend on the assumption that measurements can be chosen freely, in the same way that Bell assumed an experimenter may choose the directions of measurement a and b freely. Theories that discard this assumption of free will are termed 'superdeterministic', and are generally controversial (see for example Shimony et al. 1976 and Å. Larsson 2014). Notable proponents include Gerard 't Hooft (2012).

It now seems that local hidden variable theories cannot in any way reproduce the probabilities of the quantum mechanical formalism without invoking some notion of superdeterminism. It is clear that Jaynes was not aware of the work by Colbeck and Renner, since it was published more than a decade after his death in 1998. However, one of Jaynes's last works titled 'Probability in Quantum Theory', published in 1996, may contain a way to salvage his viewpoint of quantum determinism. Hints of this are already present in his

1989 paper, in which he noted that QM seemed to include some ontic elements, despite it often giving predictions in terms of probabilities (which are to Jaynes, after all, purely epistemic). The following section elaborates on this idea.

5.4.3 Quantum mechanics as inference

Despite the result by Colbeck and Renner (2011) there may still exist an avenue for Jaynes's interpretation of quantum mechanics as a theory of inference to be relevant. This avenue rests on the realisation that the previous arguments against local hidden variable theories relied on the notion that such theories should reproduce QM probabilities. Indeed, this is a very reasonable assumption, since the predictions made by the formalism are accurately reproduced by experiment.

Then, to see where this assumption might go wrong, consider again the viewpoint of what Ballentine has named the statistical interpretation of quantum mechanics; the state vector ψ describes a statistical ensemble of identical systems. This bears striking similarities to the original interpretation of statistical mechanics, which according to Jaynes is fundamentally a theory of inference. As such, it may be possible to interpret the quantum mechanical formalism as a theory of inference as well. While such an interpretation clearly requires local hidden variables - similarly to how statistical mechanics requires microscopic laws for its probabilities - this does not circumvent the results of Colbeck and Renner. Interestingly, the main difficulty with interpreting the QM formalism as inference may also provide the solution to this problem. In his 1996 paper, 'Probability in Quantum Theory', Jaynes notes:

> *"(...) our present QM formalism is not purely epistemological; it is a peculiar mixture describing in part realities of Nature, in part incomplete human information about Nature - all scrambled up by Heisenberg and Bohr into an omelette that nobody has seen how to unscramble. Yet we think that the unscrambling is a prerequisite for any further advance in basic physical theory. For, if we cannot separate the subjective and objective aspects of the formalism, we cannot know what we are talking about; it is just that simple. So we want to speculate on the proper tools to do this."*

As one would expect by this point, the proper tool for incorporating human information into science according to Jaynes is probability theory. However, the situation in quantum mechanics is not so simple as that in statistical mechanics; one cannot simply reinterpret QM probabilities as epistemic. The reason for this is that the probabilities of quantum mechanics do not contain all relevant information that is contained in the pure state wavefunction ψ. From Jaynes's perspective this is a hint that quantum probabilities are not what a local hidden variable theory should reproduce after all.

This lack of some information in QM probabilities is easily illustrated. Jaynes starts by expanding ψ in the energy representation $\psi(x,t) = \sum a_n(t)u_n(x)$ and notes that the physical situation this corresponds to cannot be fully described by statements such as 'the system is in state $u_1(x)$ with probability $p_1 = |a_1|^2$.' Such a description suffices to characterise quantities diagonal in the $\{u_n\}$ basis, but gives no information of relative phases, which matter for quantities not diagonal in this basis. An example is that the relative phases of degenerate energy states of an atom determine the polarisation of its resonance radiation. This is an experimentally verifiable fact, so these phases must contain some physical information (Jaynes 1996b). Thus, if one were to reinterpret these probabilities in the same way as in classical statistical mechanics, they would find that their experimental results must depend on the chosen representation of ψ; an untenable conclusion.

An epistemic quantum theory then must contain the information contained in phases in its probabilities to allow for inference based on all the relevant information. Since it is possible to know ψ, it seems that current QM probabilities throw away information.

For this reason, one cannot simply adopt the classical interpretation that a system is in state u_1 or u_2 as if they are mutually exclusive possibilities, with the suggestion that the associated probabilities are only an expression of the experimenter's ignorance. Jaynes notes that in some sense the system must be in both states simultaneously, and calls this *"the conceptually disturbing, but experimentally required, function of the superposition principle."* (Jaynes 1996b).

Jaynes conjectures that this is also what hindered Bohr from making ontic statements. Indeed, Jaynes characterises claims such as '$|a_n|^2$ is the probability that a system is in the n'th state' as unjustified, and rather supports the more epistemic statement that '$|a_n|^2$ is the probability that, if we measure the system's energy, we shall find the value corresponding to the n'th state' (Jaynes 1996b).

In a sense, it seems that ψ contains both epistemic probability and ontic elements; these different aspects will have to be unscrambled in order to reduce QM to a theory of inference. Jaynes muses that we require *"a different formalism, isomorphic in some sense but based on different variables; it was only through some weird mathematical accident that it was possible to find a variable which scrambles them up in the present way."* (Jaynes 1989).

Accordingly, he suggests that the probabilities $|\psi|^2$ are not the probabilities of inference that he is looking for; that is, the experimenter's information must be represented in some deeper hypothesis space that contains both phases and amplitudes, such that their probabilities also contain relevant phase information. Jaynes compares the system described above with a classical vibrating bell that is in a linear combination of two vibration modes with a definite relative phase, and notes that there is nothing conceptually disturbing about this picture. Of course, in this example the squared amplitude is interpreted as the energy of the mode instead of its probability, which may be another hint that the interpretation of $|\psi|^2$ as probability is causing the difficulty.

It may be interesting to note that this deeper hypothesis space essentially represents what would be the microstate of QM from Jaynes's perspective, knowledge of which would render the use of probabilities unnecessary for quantum mechanics itself. We clarify that this view operates on a distinct level from quantum statistical mechanics, where one's density matrix represents a 'classical uncertainty' - in the sense of the inference in classical statistical mechanics - in dealing with certain (pure) states that themselves contain 'quantum uncertainty' - in the sense of the above discussion on ψ.

One may perhaps consider a pure quantum state the microstate of a quantum statistical ensemble described by a density matrix - though such an interpretation faces its own difficulties (Nathaniel 2012) - yet our discussion is not concerned with this. Rather, we are concerned with the possibility of describing the uncertainty in pure states in terms of inference from incomplete information of some deeper level of reality.

Certainly, some realist theories already exist based on non-local hidden variables (Bohm 1952) or local hidden variables under the assumption of superdeterminism ('t Hooft 2012). Jaynes's interpretation however has primarily attempted to circumvent Bell's no-go theorem by pointing out a possible error in the theorem's manipulation of probabilities.

However, works by Gill, Gull, Colbeck and Renner - the latter two unbeknownst to Jaynes - have provided similar constraints on local hidden variable theories. As such Jaynes has been lead to suggest that the probabilities of QM are not pure probabilities of inference,

but rather some combination of epistemic and ontic elements that has come about by mathematical coincidence (Jaynes 1989).

Of course, one may wonder whether the notion of determinism implicit in this deeper hypothesis space is distinct from that of superdeterminism. If there is no such difference, then this idea by Jaynes may represent simply another design for some superdeterministic theory of quantum mechanics.

It is noted by Jaynes that his novel interpretation - of not viewing QM probabilities as probabilities at all - throws off a host of past difficulties with constructing deterministic theories for quantum mechanics, most notably those presented by Bell (and later Gull, Gill, and possibly Colbeck and Renner). It seems impossible to get the standard QM probabilities out of such a local hidden variable theory, but if those are not the basic probabilities after all then this is no surprise. Indeed, the failure of causal theories to reproduce the values of 'probabilities' as probabilities may perhaps then be seen as a merit of the theory; it is in this way that the original assumption of Colbeck and Renner might fail.

Jaynes explicitly notes that finding this deeper hypothesis space is work for the future, and continues his 1996 paper exploring areas of quantum theory where his interpretation might solve long-standing difficulties. Nevertheless, not much has been written on the subject since Jaynes's death in 1998, and to the author's knowledge the scientific community is not much closer to a consensus on interpretations of phenomena such as wavefunction collapse. At the current time it is not clear to the author whether Colbeck and Renner's argument excludes this deeper hypothesis space; this may be a point of further research. While any number of features in Jaynes's perspective may in the end turn out to be incorrect, his views may nonetheless represent interesting research opportunities for the future.

In closing of these considerations on quantum mechanics, let us reiterate once more the purpose of this endeavour in Jaynes's own words (Jaynes 1996b):

> *"(...) we would like to see QM as the process of making the best predictions possible from the partial information that we have when we know ψ . If we could either succeed in this, or prove that it is impossible, we would know far more about the basis of our present theory and about future possibilities than we do today."*

Chapter 6

Discussion

In this thesis we have attempted to exhaustively explore the Cox-Jaynes (CJ) approach as a justification for objective Bayesian probability theory. Furthermore, we have discussed its extension to probability theory as logic as espoused by Jaynes, and some of the latter's consequences for interpretations of physical theories.

More specifically, we began with a discussion of Bayesian methods and philosophy in Chapter 2, on which we based all that followed. Especially important for our purposes was the objective Bayesian approach to probability, which states that a probability is a unique measure of one's reasonable belief, fully determined by the available information.

A justification for this view has been provided by Cox's theorem, discussed in Chapter 3. Cox's theorem proposes some seemingly reasonable properties of a measure of reasonable belief, and proves that this measure must be a probability distribution by deriving the basic rules of probability theory from these properties. Cox's original work was extended by Jaynes, which has earned this justification the moniker of the Cox-Jaynes approach.

As we saw in Chapter 4 however, the axioms of the CJ approach are not without controversy. The density axiom in particular - originally by Paris, though already implicitly present in Cox's work - has been discussed thoroughly in the literature. While its necessity for Cox's theorem has not been questioned, its validity often has, as it - according to many, unduly - restricts the domain of the CJ approach to infinite sets of propositions. Arguments both in favour of and against this consequence of Paris's axiom exist, as do slight variations of the axiom that suffice for Cox's theorem. Nevertheless, the controversy has remained, as such variations have retained this restriction.

A solution may have been found by Terenin and Draper, who have proposed a reaxiomatisation of the CJ approach, replacing Paris's axiom with their own axioms of 'sequential continuity' and 'comparative extendability'. This approach lacks the issues of earlier versions, and derives a probability function that explicitly satisfies countable additivity, where previous attempts only explicitly assured a finite equivalent. Using Stone's representation theorem, Terenin and Draper have furthermore reconciled the Cox-Jaynes axiomatisation to Kolmogorov's original probability axioms, providing a connection between Bayesian probability and this fundament of mathematical probability.

However, the most controversial aspect of Cox's theorem - one that remains unsolved - is the validity of the representation axiom. While it may seem natural to represent reasonable belief with a single real number, we have seen that such an approach encounters problems in situations where little information is available. Appendix A elaborates on some of the current methods for constructing objective Bayesian priors in such situations,

while Appendix B further illustrates general problems with a one-dimensional theory of reasonable belief.

Although the objective Bayesian view is faced with such difficulties, Cox's theorem has proven that probability theory is the unique one-dimensional representation of reasonable belief. Jaynes built on this idea, constructing his philosophy of probability theory as logic: the idea that probability theory may be considered an extension of logic to situations in which only incomplete information is available, such that pure deduction is impossible.

This in turn lead Jaynes to the idea of the mind projection fallacy, to which Chapter 5 is dedicated. This fallacy amounts to a confusion of epistemic and ontic properties of any concept. Following Jaynes's contention that its primary appearance is in the framework of probabilities, our focus has been on the mind projection fallacy in probability interpretations as well.

Particular applications of probability theory to hypothetical experiments have shown the epistemic nature of classical probabilities and of randomisation, corroborating Jaynes's assertions and supporting probability theory as logic. Applications to the physical theory of statistical mechanics have furthermore revealed the latter's character as a theory of inference from the incomplete information represented by macrostates (Appendix A.2). This interpretation has been shown to solve a number philosophical issues within the physics of statistics, and provides an alternative understanding of the second law of thermodynamics.

Finally, we have discussed the consequences of probability theory as logic for local hidden variable theories underlying quantum mechanics (QM). It has been suggested that the historical disagreement between Einstein and Bohr on this subject primarily resulted from them speaking at disparate philosophical levels, and that Einstein's position has been uncharitably characterised later. The resolution by Bell's no-go theorem and subsequent experiments has furthermore been proposed to not be so final as is often supposed, leaving room for local deterministic theories.

Works of later researchers that restrict such theories have been briefly discussed, but it remains unclear to the author whether Einstein's and Jaynes's hopes for local deterministic theories have been conclusively ended. The reduction of quantum mechanics to a theory of inference is not so straightforward as in statistical mechanics, due to quantum probabilities not reflecting all information contained in wavefunctions from which they are derived, suggesting that our current QM formalism throws away information in calculating probabilities.

However, this difficulty simultaneously provides an avenue for a possible local deterministic theory underlying QM, as it suggests that the probabilities provided by QM may not be the probabilities that inference from an underlying hidden variable theory should precisely reproduce. The conclusion then, may be that there exists a deeper hypothesis space that does represent all of the experimenter's information, inference on which produces epistemic probabilities that corroborate experiment.

Currently, Jaynes's arguments applying the mind projection fallacy to quantum mechanics seem somewhat compelling, but not yet convincing. On the other hand, it seems difficult to find fault with his interpretation of classical statistical mechanics. Finally, the overall applicability of objective Bayesian probability as the unique measure of reasonable belief may be questioned, although Cox's theorem proves it to be the best one-dimensional option.

Certainly, further research is required to conclusively settle both the controversy surrounding the representation axiom of Cox's theorem and the validity of quantum mechanics as

a theory of inference. It seems feasible that such endeavours will produce relevant results in statistics, decision theory, philosophy and physics, though it is as of yet unclear what form those will take. In particular, a single theory of reasonable belief as well as a further understanding of quantum mechanical occurrences - such as wavefunction collapse - may be attainable.

Acknowledgements

I would like to thank my supervisor, Peter Grünwald, for finding the time in his busy schedule to supervise me during this project. His knowledge of the field of Bayesian probability is what enabled me to write on this deeply interesting subject after my initial project crashed in December.

Further thanks go out to my examiners: Ben Freivogel - who correctly pointed me towards the CWI for a project suited to my interests - and Christoph Weniger - whose enthusiasm and quick responses eased my work.

I feel that I should also thank Edwin Jaynes, not in the least for writing nearly a sixth of the literature I have used in writing this thesis. He has been likened to a thousand year old vampire before, and the justifications for this comparison have over the last few months become more apparent. So thank you, Mr Jaynes, for giving me - and hopefully others - a lot to think about.

Finally, thanks to my parents, partly because my father has been facetiously insisting on being mentioned in these acknowledgements, but more so because without them I would not be in a position to write this in the first place.

Appendix A

Methods for determining objective priors

As was mentioned in section 2.3, the objective Bayesian perspective has access to a wealth of rules for determining objective priors. This appendix serves as a more in-depth discussion of some of these rules. Specifically, the work of and philosophy of Harold Jeffreys (of the famous Jeffreys's rules) is discussed in section A.1, while Edwin Jaynes's methods of entropy maximisation and prior transformations are respectively discussed in sections A.2 and A.3. We note that the former of Jaynes's techniques may be especially interesting for those familiar with the field of statistical mechanics.

A popular method for construction objective priors that will not be treated here is that of reference analysis, for which we refer to Bernardo (1979) and Bernardo and Berger (1991). A more comprehensive overview of objective prior methods is provided by Kass and Wassermann (1996) and Ghosh (2011).

A.1 Jeffreys's rules

Harold Jeffreys is often credited with saving the objective Bayesian perspective from becoming a feature of the past in the first half of the twentieth century (Robert et al. 2009). His book *Theory of Probability* (1939) represents a comprehensive treatment of Bayesian inference from an objective Bayesian perspective, and may have been a counter to the severe criticism on the Bayesian viewpoint at the time by Fisher (see Aldrich 2008 for a extensive discussion of Fisher's criticisms). Jaynes, to date perhaps the objective Bayesian perspective's most strident proponent, has dedicated his 2003 book to Jeffreys, writing: "Dedicated to the memory of Sir Harold Jeffreys, who saw the truth and preserved it."

This section is similarly dedicated to Jeffreys. His particular Bayesian view is explored in section A.1.1, while some of his quantitative rules for constructing objective priors in problems of parameter estimation are presented in section A.1.2.

A.1.1 Jeffreys's philosophy

Jeffreys believed in the existence of an 'initial' state of knowledge (now identified as the prior), and considered it important that inferences could be made from collected data at this stage. On page 33 of the 1961 edition of his book this stage is described as one in

which an experimenter has no opinion whether a certain hypothesis is true or not. He then writes:

> *"if there is no reason to believe on hypothesis rather than another, the probabilities are equal (...) if we do not take the prior probabilities equal we are expressing confidence in one rather than another before the data are available (...) and this must be done only from definite reason."*

Thus it is clear that he subscribed to Laplace's principle of indifference in the specific case of total ignorance. Furthermore, to Jeffreys probabilities were logical relations, independent of personal belief. In the first edition of his book (Jeffreys 1931, p.10) it is written that

> *"Logical demonstration is right or wrong as a matter of the logic itself, and is not a matter for personal judgment. We say the same about probability. On a given set of data p we say that a proposition q has in relation to these data one and only one probability. If any personal assigns a different probability, he is simply wrong, and for the same reasons we assign in the case of logical judgments."*

which qualifies Jeffreys as an objectivist in the general case as well. However, it has been pointed out by Kass and Wasserman (1996, p.2) that Jeffreys's perspective evolved over time to the more nuanced view that there may be multiple proper ways to determine objective priors, so that scientific convention should resolve any remaining ambiguity. In this he compared the choice of 'objective' prior to that of a system of units.

Indeed, in later editions of *Theory of Probability* the latter of the two passages quoted is absent. Here Jeffreys took a reasonable degree of probability as the primitive concept as opposed to the uniquely determined logical relations he advocated earlier. In his own words (p.22 of *Theory of Probability*, 1957 edition) *"if we like, there is no harm in saying that probability expresses a reasonable degree of belief"*, and (Jeffreys 1955):

> *"It may still turn out that there are many equally good methods (...) if this happens there need be no great difficulty. Once the alternatives are stated clearly a distinction can be made by international agreement, just as it has been in the choice of units of measurement and many other standards of reference."*

Thus the later Jeffreys did not insist on a particular representation of ignorance, rather presenting the principles he developed as a guide that in some cases might yield unique such a unique representation. In other cases some sort of convention would be necessary to handle inference problems consistently (Kass and Wasserman 1996, p.3). Some of Jeffreys's quantitative rules for constructing priors are discussed in the following section.

A.1.2 Specific and general rules

As one may have gathered from the previous section, Jeffreys's methods are less rigorous than some may wish. One consequence of this is the importance of convention in his methods, as was noted above. Jeffreys proposed different rules in differing situations, and notably made a distinction between problems of parameter estimation and those of hypothesis testing. Here we will briefly describe the former, and refer to Kass and Wasserman (1996, section 2.3) for a discussion of the latter.

Jeffreys's earliest approach was to base his ignorance priors on mathematical properties of the parameter in question. For example, consider estimating some quantity v, which under the available knowledge is limited by a continuous range of V values. This set V

can be parametrised by some smooth function $f(v) = \sigma \in \Sigma$. Now σ is the parameter of interest and Σ the associated parameter space that contains the possible values. The trivial example is the function f that maps v onto its dimensionless magnitude. The first step in Bayesian estimation must now be to represent this initial knowledge (that $\sigma \in \Sigma$) about v in a probability function.

Jeffreys proposed basing the prior on the specific form of parameter space under consideration. For instance, for the full real interval $(-\infty, \infty)$ he suggested a constant (uniform) prior density, because such a prior is invariant under linear transformations (Seidenfeld 1979). Such arguments are often at the basis of Jeffreys's rules; he argued that consistency of the prior over important families of alternative parametrisations was an important quality. Indeed, one should not wish to use priors that change under parametrisations that leave the parameter space invariant, since those are surely not objective.

Note that using this rule for the interval $(-\infty, \infty)$ will lead to an improper prior - one that does not integrate - as $\int_{-\infty}^{\infty} c\, d\sigma \to \infty$ for all constant c. Jeffreys did not consider this to raise any fundamental difficulties (Kass and Wasserman 1996, p.3).

For positive intervals, most notably $(0, \infty)$, Jeffreys proposed the prior $\pi_\sigma(\sigma) = \frac{1}{\sigma}$, primarily because it has the property of being invariant under power transformations. Indeed, power transformations map the interval $(0, \infty)$ to itself, and one would not want such a transformation to affect the prior density.

This invariance be understood from the following argument: by preservation of probability $\pi_\sigma(\sigma)d\sigma = \pi_\gamma(\gamma)d\gamma$, from which it follows that $\pi_\sigma(\sigma) = \pi_\gamma(\gamma)\frac{d\gamma}{d\sigma}$. Then, if $\gamma = \sigma^n$, we have $\pi_\sigma(\sigma) = \pi_\gamma(\gamma)\, n\, \sigma^{n-1} = \frac{n}{\sigma}$ up to a normalisation factor.

Note that the two rules mentioned above are moreover mutually consistent under a transformation of the parameter space from $(0, \infty)$ to $(-\infty, \infty)$ by $\theta \to \ln(\theta)$.

Jeffreys later proposed his more general rule based on the Fisher information matrix $I(\sigma)$ (Jeffreys 1946). The prior is

$$\pi_\sigma(\sigma) \propto \det\left(I(\sigma)\right)^{\frac{1}{2}} \tag{A.1}$$

with

$$I(\sigma)_{ij} = \mathbb{E}\left(-\frac{\partial^2 l}{\partial\sigma_i\, \partial\sigma_j}\right) \tag{A.2}$$

and l the log-likelihood. Using any other parametrisation γ gives

$$\pi_\sigma(\sigma_i) \propto \det\left(\mathbb{E}\left(\frac{\partial^2 l}{\partial\gamma_k\, \partial\gamma_m}\frac{\partial\gamma_k}{\partial\sigma_i}\frac{\partial\gamma_m}{\partial\sigma_j}\right)\right)^{\frac{1}{2}} \tag{A.3}$$

$$\propto \det\left(\mathbb{E}\left(\frac{\partial^2 l}{\partial\gamma_k\, \partial\gamma_m}\right)\right)^{\frac{1}{2}} \det\left(\frac{\partial\gamma_k}{\partial\sigma_i}\frac{\partial\gamma_m}{\partial\sigma_j}\right)^{\frac{1}{2}} \tag{A.4}$$

$$\propto \pi_\gamma(\gamma_k) \det\left(\frac{\partial\gamma_k}{\partial\sigma_i}\right) \tag{A.5}$$

which is the multidimensional equivalent of the earlier rule, so that this general rule is invariant under power transformations as well.

In the 1961 edition of his book Jeffreys notes that this general rule may conflict with the previously stated rules in some cases. As a specific example, for data distributed as $N(\mu, \sigma^2)$, the general rule will give $\pi(\mu, \sigma) = \frac{1}{\sigma^2}$, while the previous rule (on the $(0, \infty)$ interval) gave $\pi(\mu, \sigma) = \frac{1}{\sigma}$ (Kass and Wasserman 1996). This problem is solved by considering μ and σ separately a priori; when applying the general rule to either parameter, the other should be held fixed. This procedure leads to the desired prior $\pi(\sigma) \propto \frac{1}{\sigma}$. Jeffreys proposed this alteration to his rule for general problems involving both location and scale parameters.

Jeffreys's rules have other problems too. For instance, if V is the unit interval $(0, 1)$ it is unclear whether the positive parameter rule $\sigma = m(1 - m)$ or the uniform distribution rule should be used. Furthermore, if we learn that a parameter we considered real-valued is actually limited to the positive reals, is the rule then to limit the uniform distribution at the origin, or should some form of the positive parameter rule $1/\sigma$ be used?

Jeffreys's original response to these criticisms was that the dimensions associated with v give information for deciding on the relevant family of reparametrisations, and hence on the correct ignorance prior. In many practical applications v is a physical quantity, not just a number, and so this approach may have merit. Indeed, the convincing strength of Jeffreys's rules was their ability to reproduce classic statistical results in a Bayesian manner (Seidenfeld 1979), so his methods have enjoyed success.

A later innovation by Jeffreys - the theory of invariants (Seidenfeld 1979) - was to not just consider v, but also the distribution of the random variable whose testing will provide the statistical information about v. In this way, v can typically be translated into direct probabilities of the model under consideration. Is is then the invariant parametrisations of the model that determine which family of transformations is important for the quantity of interest v.

For example, v may be the position of a bent coin's centre of mass. It is not immediately clear how to parametrise this, nor what the relevant transformations should be. However, a method for gaining information about v may be simply flipping the coin; a process that is described by a binomial distribution. If the binomial parameter is found to be a half, then this may correspond to the statement that the coin is fair, and thus that the centre of mass is the geometric centre of the coin (Seidenfeld 1979). In this way it has become possible to find a relation between the binomial parameter of the experimental model and v. It may then be concluded that the relevant family of parametrisations should be the one that leaves the binomial distribution invariant.

For more on this theory of invariants we refer to Huzurbazar (1976). This furthermore concludes our brief overview of Jeffreys's rules. We refer again to Kass and Wasserman (1996) for a more comprehensive discussion.

A.2 The principle of maximum entropy

The principle of maximum entropy (MaxEnt) was first introduced by Jaynes in a set of two papers titled 'Information Theory and Statistical Mechanics I & II' (Jaynes 1957a; Jaynes 1957b). To both information theorists and physicists alike, these titles may seem rather vague: it is not immediately clear that there is (or ought to be) a connection between the fields of information theory and statistical mechanics. Nevertheless, to those familiar with both fields, there is a clear similarity in that both fields make heavy use of the concept of entropy.

This connection is first elaborated on in section A.2.1, which also serves as an intuitive introduction to the principle of entropy maximisation. Afterwards, in section A.2.2 it is argued that statistical mechanics may be considered an application of Jaynes's principle by using it to derive the microcanonical ensemble. Section A.2.3 briefly discusses applications to the density matrix formalism, often used to describe mixed systems, which are more elaborately discussed in the second of Jaynes's aforementioned papers. Finally, we close with a discussion of applications to continuous probability distributions in section A.2.4. This discussion then naturally leads into the next and final method of prior selection we shall discuss; the prior transformation groups of section A.3.

A.2.1 Information theory and statistical mechanics

The connection between information theory and statistical mechanics is most apparent by their mutual use of entropy. In statistical thermodynamics, the Boltzmann entropy is a measure the number of microstates corresponding to a given macrostate in thermodynamic equilibrium. The equation is $S_B = k_B \ln \Omega$, where S_B is the Boltzmann entropy, k_B the Boltzmann constant, and Ω the multiplicity of a given macrostate (i.e. the number of microstates). It has been shown that this equation is equivalent to the classical heat engine entropy $\delta S = \delta Q/T$ (Jaynes 1965).

Entropy is often considered equivalent with terms such as 'chaos' and 'disorder', but this tends to lead to confusion. For instance, a glass of water might intuitively be considered more ordered than one of crushed ice, yet the latter has the lower entropy. A better description of entropy then, is that it is a measure of information. This may be interpreted in two ways: one, it is the information in the system being considered; two, it is the lack of information one has about the system.

Consider then a more advanced definition of entropy, the classical Gibbs entropy $S_G = -k_B \sum_i p_i \ln p_i$. Here p_i represents the probability that a system in a given macrostate is in a microstate i. The Gibbs entropy is a generalisation of the Boltzmann entropy for systems out of equilibrium, where microstates of a system are not equally probable. Indeed, setting all p_i equal will yield the result that $S_G = S_B$. Note that this is the maximum value of S_G for a given macrostate, and as such an isolated system will tend to this configuration by the second law of thermodynamics. This shows that 'thermodynamic equilibrium' and 'maximum Gibbs entropy' are equivalent conditions.

In this definition of entropy, the connection with information is more apparent. From a Bayesian perspective, the probabilities p_i are statements about one's knowledge - or rather, lack of knowledge - about the underlying microstates. When all these probabilities are equal, the Gibbs entropy is maximal and one's knowledge minimal, since it is impossible to reasonably select one microstate over another. Were one to select a microstate i as more plausible in this case, then that would constitute a biased choice, which should be avoided in statistical inference. This is reminiscent of the principle of indifference from section 4.1. It stated that equal probabilities should be assigned in the absence of knowledge about propositions; any other assignment would be biased. This principle is now motivated by a reason other than not knowing any better; the principle of indifference naturally follows from an unconditional maximisation of entropy. In the words of Jaynes (1957a):

> *"The maximum-entropy distribution may be asserted for the positive reason that it is uniquely determined as the one which is maximally noncommittal with regard to missing information, instead of the negative one that there was no reason to think otherwise."*

One might now wonder why entropy specifically is an appropriate measure of information. This is due to Shannon's (1948) monumental work that is often considered the birth of the field of information theory. In it, Shannon showed that the (Shannon) entropy $H = -\sum_i p_i \log p_i$ with p_i a discrete probability distribution is a unique measure for uncertainty that satisfies all the 'conditions which make it reasonable' (Jaynes 1957a). Note that this is equivalent to the definition of the Gibbs entropy, barring a factor k_B and a different base logarithm. The addition of the Boltzmann constant merely changes the units in which entropy is measured, while the logarithm represents a scale transformation from nats to bits.

Of course, even with this definition the notion that entropy is a measure of ignorance remains conceptually vague. To lessen this sense of vagueness let us consider the questions of 'whose ignorance?' and 'about what?'. The answer to the first is always 'the experimenter', i.e. the one considering the system. Indeed, as we shall see later entropy is a tool for inference from the Bayesian perspective, and hence its value lies in describing the state of knowledge of the one inferring. The second question has a similarly trivial answer, namely 'the system under consideration', but systems are usually part of larger systems that then also require description. A more useful way to think of entropy is perhaps as a property that is uniquely determined provided that it is specified which degrees of freedom in the universe are within the system (and which are not).

With the intuitive rationale behind the maximum entropy principle introduced, we consider an example of its application in the next section. We will attempt to derive the microcanonical ensemble of statistical physics from a few basic assumptions together with this principle.

A.2.2 Statistical mechanics as logical inference

In his two papers, Jaynes argues that statistical mechanics is simply an application of logical inference and Shannon information theory. In particular, he supports this by deriving the microcanonical ensemble via maximisation of a system's entropy using only knowledge of the system's energy. Let us follow this derivation to elucidate the process.

A quantity x can take a number of discrete values x_i, $i = 1, 2, \ldots, n$. The corresponding probabilities p_i are unknown, but the expectation value of a function $f(x)$ given by $\langle f(x) \rangle = \sum_i^n p_i f(x_i)$ is known. Here x may be identified as a microstate, and $\langle f(x) \rangle$ as the energy of a given system. The goal is to estimate the value of another function $g(x)$, which corresponds to another macroscopic property (for instance $\langle g(x) \rangle$ may represent the system's temperature).

Thus, we have a system of n variables with two known equations, the other being the normalisation condition $\sum_i p_i = 1$. This seems unsolvable, as generally $(n-2)$ more function are required to solve such a system. Yet, this in particular is a problem of specifying probabilities, so one might consider using the principle of indifference. Such an approach will fail in general however, since it requires setting all probabilities equal, which is only correct for a system in thermodynamic equilibrium. Indeed, note that the principle of indifference requires no knowledge of the system at all, while in this problem some knowledge is available.

As remarked above, the principle of indifference following from maximising entropy in situations of complete ignorance. The question is then how to apply this principle while still using the available knowledge. Jaynes notes the similarity to a constraint problem: one should maximise entropy subject to the given constraints; that is, subject to the

available knowledge. The problem then reduces to a simple maximisation with Laplace multipliers. One maximises the function

$$L = -\sum_i^n p_i \ln p_i + \lambda \left(\sum_i^n p_i f_i - \langle f \rangle \right) + \mu \left(\sum_i^n p_i - 1 \right) \tag{A.6}$$

with respect to the discrete distribution p_i, where the first term is the Shannon entropy (now in nats instead of bits), and λ, μ are the Lagrange multipliers. This yields:

$$\frac{\partial L}{\partial p_i} = -\ln p_i - 1 + \lambda f_i + \mu = 0. \tag{A.7}$$

It follows that $p_i = \exp(\lambda f_i + \mu - 1)$. Plugging this into the normalisation condition for p_i we obtain

$$\begin{aligned} 1 &= \sum_i p_i \\ &= \sum_i \exp(\lambda f_i + \mu - 1) \\ &= \exp(\mu - 1) \sum_i \exp(\lambda f_i) \end{aligned} \tag{A.8}$$

or $\sum_i \exp[\lambda f_i] = \exp(1 - \mu)$. Since the Lagrange multiplier μ originates from the normalisation condition on p_i, one might identify $\exp(\mu - 1)$ as a normalising factor. Call this Z in anticipation of obtaining the partition function of the microcanonical ensemble. This gives

$$Z = \exp(\mu - 1) = \frac{1}{\sum_i \exp(\lambda f_i)} \tag{A.9}$$

such that

$$p_i = \frac{1}{Z} \exp(\lambda f_i) = \frac{\exp(\lambda f_i)}{\sum_i \exp(\lambda f_i)}. \tag{A.10}$$

On the other hand, the condition for the expectation value of $f(x)$ now states:

$$\begin{aligned} \langle f \rangle &= \sum_i p_i f_i \\ &= \sum_i f_i \exp(\lambda f_i + \mu - 1) \\ &= \frac{\partial}{\partial \lambda} \ln Z \end{aligned} \tag{A.11}$$

Identifying f with energy E allows for the identification of λ with $\frac{1}{k_B T}$. Thus the microcanonical ensemble has been obtained by statistical inference alone, relying on MaxEnt. Jaynes goes on to repeat this procedure for more complicated systems, including the grand

canonical ensemble, where the essential difference is that the number of molecules n_s of a type s is no longer considered fixed, but also given in terms of expectation $\langle n_s \rangle$.

In this way, Jaynes argues for the principle of indifference via the maximum entropy principle. Indeed, MaxEnt may be considered an extension of indifference, as it naturally handles cases in which some information is available, rather than none. Jaynes emphasises the role of statistical inference in solving physical problems in an unbiased manner:

> *"In the problem of prediction, the maximization of entropy is not an application of a law of physics, but merely a method of reasoning which ensures that no unconscious arbitrary assumptions have been introduced."*

It is important to note that there is no claim that the predictions made by this process of reasoning from incomplete information must be 'right'; only that they are the best that can be made from the available information. As Jaynes later remarks, this should come as no surprise, and this is the most that any science could ever have pretended to do (Jaynes 1996b).

The maximum entropy principle has applications outside of statistical mechanics precisely because it is merely a theory of inference. Examples include spectrum analysis, image reconstruction, crystallographic structure determination, econometrics, and indeed any problem with the structure of a distribution over hypotheses and some extra information that can be modelled as a constraint on that distribution (Jaynes 1984a).

The application to statistical mechanics presented here was only chosen because it is the most obvious, as Shannon's measure for information - the entropy - was already familiar to physics under a different interpretation. The success of MaxEnt in statistical mechanics is the first hint that statistical mechanics is not a theory of physics at all, but rather a theory of inference. This idea was discussed in more detail in Chapter 5, particularly section 5.3.

A.2.3 Application in the density matrix formalism

Much more can be said about the validity, usefulness and interpretation of the maximum entropy principle. For a more complete discussion we refer to Jaynes's first paper on the subject 'Information Theory and Statistical Mechanics I', particularly section four. His second paper 'Information Theory and Statistical Mechanics II' deals with systems that evolve in time using the density matrix formalism. It is thus worth mentioning the connection between that formalism and the classical definition of entropy.

For a mixed system described by a density matrix ρ, the Von Neumann entropy S_N is defined as $S_N = \mathrm{Tr}\,(\rho \ln \rho)$ (the constant k_B is sometimes included in this definition as well) (Porter 2004). Since the trace is a basis independent operation, we may opt to work in the basis in which ρ is diagonal. Then the Von Neumann entropy reduces to the trace of a diagonal matrix $(\rho_i \ln \rho_i)$, where ρ_i is the i'th eigenvalue of ρ. Since the trace of a matrix is the sum of its diagonal, we obtain the expression $S_N = \sum_i \rho_i \ln \rho_i$. This reduces to the Gibbs entropy by noting that the i'th eigenvalue of a system's density matrix is the probability p_i that the system is in the pure state i.

We remark that the use of the terminology 'the probability that the system is in state i' should not be taken literally. From the Bayesian perspective, such probabilities are representative of one's ignorance of the system, resulting from a lack of knowledge of the underlying causes, and not a property of the system itself. This is the perspective that Jaynes takes in his papers, as he discusses in section two of his first paper on maximum entropy (Jaynes 1957a).

A.2.4 Application to continuous probability distributions

The applications provided until now have only considered discrete probability distributions, but the principle of entropy maximisation is certainly applicable to continuous distributions as well. However, some more work is required to make sense of entropy in such a situation; let us examine here why this is the case.

Shannon (1948) originally gave the formula for continuous entropy (differential entropy) as

$$H = -\int p(x) \ln p(x) dx \tag{A.12}$$

with $p(x)$ the probability density function. This was, however, not the result of any derivation; instead, Shannon simply replaced the discrete parts of his entropy formula with continuous parts. This expression turned out to lack some useful properties his original expression possessed, primarily since probability densities can be larger than 1. For instance, this continuous entropy could take negative values; taking $p(x) \sim \text{Uniform}(0, \frac{1}{2})$ yields

$$H = -\int_0^{\frac{1}{2}} 2 \ln 2 \, dx = -\ln 2. \tag{A.13}$$

Thus, this expression for entropy may not be useful for the purpose of MaxEnt. Furthermore, Shannon's continuous entropy is not invariant under a continuous change of variables $x \to y(x)$, which carries its own problems. To illustrate this, note that the probability in a differential area should not change under such a transformation: $p(x)dx = p(y)dy$. Then - misusing notation a bit for simplicity - it follows that

$$p(x) = p(y)\frac{dy}{dx} \tag{A.14}$$

and thus

$$\begin{aligned} \int p(x) \ln p(x) dx &= \int p(y)\frac{dy}{dx} \ln \left(p(y)\frac{dy}{dx} \right) dx \, \frac{dy}{dy} \\ &= \int p(y) \ln \left(p(y)\frac{dy}{dx} \right) dy \\ &\neq \int p(y) \ln p(y) \, dy. \end{aligned} \tag{A.15}$$

This is problematic, since it means the choice of variables changes the maximisation problem, which may lead to different probabilities. Jaynes solves this problem in a later paper by taking the limit of the discrete entropy (Jaynes 1963). This yields the following invariant expression:

$$H = -\int p(x) \ln \frac{p(x)}{m(x)} dx. \tag{A.16}$$

Here $m(x)$ is an 'invariant measure', proportional to the density of the discrete points in continuous limit. Jaynes remarks on the terminology in a later paper (Jaynes 1968):

> *"In all applications so far studied, $m(x)$ is a well-behaved continuous function, and so we continue to use the notation of Riemann integrals; we call $m(x)$ a 'measure' only to suggest the appropriate generalization, readily supplied if a practical problem should ever require it."*

With this new expression, the previous methods can be applied to continuous problems in a way that leaves the probabilities and partition function independent of the choice of variables. The measure $m(x)$ however, does not disappear from the expressions for probabilities, which means those probabilities depend on the choice of measure. To gain more insight into this shortcoming, consider the maximisation problem when no information is known, such that no Lagrange multipliers beyond that of normalisation appear in the derivation. The goal then is to maximise

$$L = -\int p(x) \ln \frac{p(x)}{m(x)} dx + \mu \left(\int p(x) dx - 1 \right) \tag{A.17}$$

with respect to $p(x)$. Thus

$$\frac{\partial L}{\partial p(x)} = -\ln \frac{p(x)}{m(x)} - 1 + \mu = 0, \tag{A.18}$$

such that

$$p(x) = m(x) \exp(\mu - 1). \tag{A.19}$$

Plugging this into the normalisation condition $\int p(x)dx = 1$ yields

$$1 = \int m(x) \exp(\mu - 1) dx = \exp(\mu - 1) \int m(x) dx, \tag{A.20}$$

such that

$$\exp(\mu - 1) = \left[\int m(x) dx \right]^{-1} \tag{A.21}$$

and therefore

$$p(x) = \left[\int m(x) dx \right]^{-1} m(x). \tag{A.22}$$

This shows that $m(x)$ is equal to the prior distribution up to a constant factor. It seems that the ambiguity that comes from the measure originates from the same problem that we have been trying to solve; how to find an objective prior distribution. The issue is now stated in the language of choosing a proper measure, but it remains an issue nonetheless.

We note that equation (A.16) is equal to the (negative) continuous version of the Kullback-Leibler divergence from m to p (Bishop 2006). This quantity can be interpreted as the information gained when the prior belief described by $m(x)$ is updated to the posterior belief $p(x)$ (Burnham and Anderson 2002). Maximising the expression in equation (A.16) is then equal to minimising this information gain.

We have already noted that $m(x)$ is similar to the total ignorance prior in the context of the maximum entropy principle. Thus, in the continuous case, MaxEnt states that one should attempt to minimise the information one garners from a set of constraints. In the context of statistical physics, for instance, this means that the principle is a natural way to ensure one does not claim to know more than the given information (usually in the form of expectation values $\langle f \rangle$) allows. Such an interpretation agrees with the motivation behind MaxEnt: the maximisation of ignorance that results in unbiased conclusions.

The minimisation of Kullback-Leibler divergence is also known under the names of 'principle of minimum discrimination information', 'principle of minimum cross-entropy', and 'minxent'. These methods may also be considered extensions of the principle of insufficient reason (Burnham and Anderson 2002).

While this digression explains the motivation behind maximising (A.16), it does not provide a method for choosing the proper measure $m(x)$ correctly in the first place. For this a different logical method is required, and Jaynes provides one by introducing prior transformation groups. A discussion of this method can be found in Appendix A.3 hereafter.

A.3 Prior transformation groups

The principle of maximum entropy, treated in Appendix A.2, provides a powerful tool for determining objective priors in many cases. However, for continuous problems the original issue of choosing a proper total ignorance prior remains, in the language of selecting a proper measure.

In order to find such a prior, one might consider whether some of its properties can be inferred. For instance, while an objective prior may change under parameter transformations, its interpretation should not change. That is, a prior $f(\sigma)$ in a coordinate system S may transform to $g(\sigma')$ in system S' under some parameter change that leaves the problem invariant, but it should be true that $f(\sigma) = g(\sigma)$, and equivalently that $f(\sigma') = g(\sigma')$, even though $f(\sigma)$ and $g(\sigma')$ will in general differ by some Jacobian factor.

It turns out that this requirement is enough to uniquely determine an ignorance prior for some problems. This is due to a method by Jaynes, which he refers to by the moniker of 'prior transformation groups'. The name was chosen because of the close connection to group theory; to a large extent, it is about finding symmetry in a problem and exploiting that symmetry to find a unique prior. Here we will illustrate the approach with an example Jaynes provides. For a more comprehensive discussion we refer to the original paper (Jaynes 1968).

Consider the situation of sampling from the two-parameter distribution

$$p(dx|\mu\sigma) = h\left(\frac{x-\mu}{\sigma}\right)\frac{dx}{\sigma} \tag{A.23}$$

where h is a nonnegative and normalised function. Note that one can infer from this distribution that μ is a position parameter and σ a scale parameter. The goal is now to estimate μ and σ given a sample $\{x_1, \ldots, x_n\}$ from this distribution. In an attempt to do this, we introduce the prior $f(\mu, \sigma)d\mu d\sigma$.

At this point we need to further specify the function $f(\mu, \sigma)$. The available information is that the prior should exhibit complete ignorance, beyond knowing the purpose of μ

(location) and σ (scale), which can be inferred from the distribution. The crucial insight by Jaynes is now the following (Jaynes 1968, p.17):

> *"If a change of scale can make the problem appear in any way different to us, then we were not completely ignorant; we must have had some kind of prior knowledge about the absolute scale of the problem. Likewise, if a shift of location can make the problem appear in any way different, then it must be that we had some kind of prior knowledge about location. In other words, complete ignorance of a location and scale parameter is a state of knowledge such that a change of scale and a shift of location does not change that state of knowledge."*

In other words, the prior $f(\mu, \sigma)$ should be invariant under transformations of position and scale. Consider then the following transformation of $(x, \mu, \sigma) \to (x', \mu', \sigma')$

$$\begin{aligned} x' - \mu' &= a(x - \mu) \\ \mu' &= \mu + b \\ \sigma' &= a\sigma \end{aligned} \tag{A.24}$$

with $0 < a < \infty$ and $-\infty < b < \infty$. Note that μ and σ transform respectively as a position and scale parameter, such that this is the most general transformation allowed. The transformation of x is fully determined by the other two and the fact that the distribution should be unchanged when expressed in the new variables:

$$p(dx'|\mu'\sigma') = h\left(\frac{x' - \mu'}{\sigma'}\right)\frac{dx'}{\sigma'}. \tag{A.25}$$

Equation (A.24) perhaps gives some indication of the importance of groups in this approach. The transformations for μ and σ can each be interpreted as forming a group, and the problem at hand should be unchanged under group operations. Under the transformation in (A.24) the prior distribution becomes

$$f(\mu, \sigma)d\mu d\sigma = f(\mu, \sigma)\frac{d\mu}{d\mu'}\,d\mu'\,\frac{d\sigma}{d\sigma'}\,d\sigma' = a^{-1}f(\mu, \sigma)d\mu' d\sigma' = g(\mu', \sigma')d\mu' d\sigma', \tag{A.26}$$

with

$$g(\mu', \sigma') = a^{-1}f(\mu, \sigma). \tag{A.27}$$

Note that we may now equivalently consider the problem of estimating μ' and σ' given a sample $\{x'_1, \ldots, x'_n\}$; the sampling distributions and priors are the same in both problems, in terms of their respective coordinates.

Furthermore, since the problems are equivalent, the same prior distribution should be assigned to both problems; indeed, the purpose of this entire endeavour is to find an objective prior, and one could hardly call a prior objective if one decides on using a different prior in an equivalent problem. By this argument then, it must be true that

$$f(\mu,\sigma)d\mu d\sigma = g(\mu,\sigma)d\mu d\sigma \tag{A.28}$$
$$f(\mu,\sigma) = g(\mu,\sigma) \tag{A.29}$$

Combining equations (A.24) and (A.27) moreover yields

$$f(\mu,\sigma) = a\, g(\mu',\sigma') = a\, g(\mu + b, a\sigma) \tag{A.30}$$

and plugging the right hand side of (A.29) into this, we obtain

$$f(\mu,\sigma) = a f(\mu + b, a\sigma). \tag{A.31}$$

Note that requiring $f(\mu',\sigma') = g(\mu',\sigma')$ would have resulted in the same outcome. Jaynes provides the general solution to this equation (Jaynes 1968, equation (54) on p.18):

$$f(\mu,\sigma) = \frac{C}{\sigma} \tag{A.32}$$

with C a constant. This then is the objective prior for this specific problem. It corresponds to Jeffreys's rule - discussed in section A.1.2 - that a uniform probability should be assigned of the form $\frac{d\mu d\sigma}{\sigma}$. Indeed, this principle of using transformation groups is fairly reminiscent of Jeffreys's use of invariants, also treated in that section.

The purpose of this example was to show that this principle of transformation groups allows for a precise definition of complete ignorance. In practice an experimenter often has access to some information, such as the model in question, or the units of some parameter, and is thus not completely ignorant in the naive sense. By specifying a set of transformations that convert the problem into an equivalent one, this initial ignorance is defined more rigorously. Then, the requirement that both the original and transformed problem should produce the same prior distribution restricts the form of the prior. With sufficient restrictions it becomes possible to uniquely define an objective prior (Jaynes 1968).

This concludes our overview of methods for constructing objective Bayesian priors. A popular method that has not been treated here is that of reference analysis, for which we refer to Bernardo (1979) and Bernardo and Berger (1991). A more comprehensive overview of objective prior methods is provided by Kass and Wassermann (1996) and Ghosh (2011).

Appendix B

Issues with ignorance

In section 4.1 we discussed the first axiom of Cox's theorem, which states that one's reasonable belief may be represented by a single real number. This axiom is controversial for a number of reasons, and although some of its opposition was shown to be unconvincing there remained a particular issue: the representation of reasonable belief in the presence of ignorance, discussed in section 4.1.2.

This issue has remained a central problem of the objective Bayesian philosophy, as evidenced by the wealth of criticism that focusses on it. In this appendix we provide a few of these arguments in order to give an idea of the difficulties. We also remind the reader that objective Bayesian philosophy - apart from the problem of ignorance - moreover faces the complication that in many situations there exist multiple seemingly valid rules for constructing objective priors. This conflicts with the fundamental objective Bayesian idea of providing unique representations of reasonable belief.

The first case discussed here is one by Cox himself, who considered the principle of indifference to be mathematically flawed. His argument will be presented in section B.1. Then, a brief discussion of imprecise probability theory - a two-dimensional theory of probability mentioned in section 4.1.2 - will be provided in section B.2, as it may provide an alternative to one-dimensional objective Bayesian theory. This will be followed by an analysis of the Ellsberg paradox, which may point to an exception in rational reasoning that is not captured by Bayesian probability. This exception is not present in imprecise probability theory, which might be considered an argument to favour it over one-dimensional probability. Finally, section B.4 treats a work by Teddy Seidenfeld, who claims Bayesian updating via conditionalisation (Bayes's theorem) is inconsistent with principles of Jeffreys's and Jaynes's, thus arguing against the objective Bayesian perspective.

B.1 Cox's argument against indifference

The principle of indifference states that, when no information relevant to a set of propositions is known, all propositions should a priori be assigned equal probability if they are mutually exclusive. This is the principle originally used by Laplace to form the priors required for solving problems with Bayesian methods (Stigler 1986). The principle has been rather controversial, not in the least because it seems only valid for problems with clear symmetry (such as a coin or die). Indeed, the maximum entropy methods discussed in section A.2 provide a generalisation for the principle of indifference, and show that it naturally arises only in cases where no relevant information is available.

Cox himself, however, disagreed with the principle even in such symmetrical cases and provided a counterargument in his 1961 paper (Cox 1961, p.29-34). The argument is one of reductio ad absurdum; he assumes the principle of indifference to be true, and then derives a result that most would find indefensible. In the following we will adapt Cox's notation to be more in line with that of Van Horn (2003).

In order to correctly invoke the principle of indifference, our states of information should contain no knowledge relevant to the propositions in question beyond that they are mutually exclusive. Considering only a proposition A, the corresponding indifferent state of information X_A may be written as $(A \vee \neg A)$, such that the only available knowledge is that A is either true or false; that is, there are two mutually exclusive possibilities. In accordance with the indifference principle then, we have $(A|X_A) = (A|A \vee \neg A) = 1/2$ and the same for $(\neg A|X_A)$. Here we follow Cox's notation of propositions in the Boolean algebra, but note that this is equivalent to probabilistic notation, i.e: $(A|B) \equiv P(A|B)$.

Cox then introduces another proposition B, for which he separately invokes the principle of indifference; $X_B = (B \vee \neg B)$, and thus $(B|X_B) = 1/2$. Now, using the identity $(A \vee \neg A) \vee C = C$ for any proposition C, we see that $(A \vee \neg A) = (A \vee \neg A) \vee (B \vee \neg B) = (B \vee \neg B)$. We then consider the combined proposition $(A \wedge B)$ and write

$$(A \wedge B|A \vee \neg A) = (A \wedge B|B \vee \neg B). \tag{B.1}$$

Using the product rule twice

$$(A \wedge B|A \vee \neg A) = (A|A \vee \neg A)(B|A, (A \vee \neg A)) = (A|A \vee \neg A)(B|A) \tag{B.2}$$
$$(A \wedge B|B \vee \neg B) = (B|B \vee \neg B)(A|B, (B \vee \neg B)) = (B|B \vee \neg B)(A|B) \tag{B.3}$$

and plugging these into (B.1), we obtain

$$(A|A \vee \neg A)(B|A) = (B|B \vee \neg B)(A|B). \tag{B.4}$$

Earlier it was concluded that both $(A|A \vee \neg A)$ and $(B|B \vee \neg B)$ were equal to $1/2$, which leads to the result that $(B|A) = (A|B)$ for any propositions A, B.

Cox calls this conclusion 'monstrous', saying that $(A|B)$ and $(B|A)$ can have any ratio from zero to infinity (Cox 1961, p.32). The solution, proposes Cox, is then to conclude that the assumption that $(A|A \vee \neg A) = 1/2$ is false. He suggests that the probabilities are undefined when the 'hypothesis' (here: state of information) is the 'truism' (here: the trivial state $(A \vee \neg A)$), except those corresponding to the truism and the absurdity themselves; that is

$$1 = (A \vee \neg A|A \vee \neg A) \tag{B.5}$$
$$0 = (\neg(A \vee \neg A)|A \vee \neg A) = (\neg A \wedge A|A \vee \neg A) \tag{B.6}$$

which are clearly true. Thus, Cox shows that applying the principle of indifference can lead to indefensible conclusions, which leads him to conclude that it is not a valid tool for statistical inference.

Of course, the conclusion that $(A|B) = (B|A)$ is not surprising considering the experimenter knows nothing of A and B, since there is no information available that allows

them to differentiate between the two plausibilities. The principle of indifference provides a reasonable answer to the question 'given that we know nothing of A and B, which of $(A|B)$ and $(B|A)$ should we find more plausible?', namely that we should find them equally plausible.

On the other hand, the definite conclusion that $(A|B) = (B|A)$ seems indefensible, given that there is no available information that indicates this conclusion. This points again to the problem with one-dimensional representations of plausibility, discussed in section 4.1.2; the inability to provide an evaluation of the quality of information. There is a pronounced difference between the conclusions '$(A|B) = (B|A)$ because there is no reason to consider them having distinct values' and '$(A|B) = (B|A)$ because there is definite information supporting this conclusion'.

Thus, while Cox's theorem proves that Bayesian probability theory is the proper one-dimensional description of plausibility, this argument indicates that a one-dimensional theory may not be sufficient. Two-dimensional theories may be required to properly describe plausibility for all purposes. The next section of this appendix briefly treats such a two-dimensional theory.

B.2 Imprecise probability theory

The existence of two-dimensional theories as an alternative to Bayesian probability theory for representing plausibilities was mentioned in section 4.1.1. The main drive behind such theories is that they are better able to represent the ignorance one might have about a proposition than the one-dimensional probability theory. This is especially apparent in cases where little information is available, such that there is large uncertainty in a subjective probability assignment.

The objective Bayesian view does not share this problem in principle, since its probabilities are uniquely determined by the available information. However, it is still not clear that there exists a method for assigning objective priors that works for all cases. Moreover two-dimensional theories may contain useful properties that one-dimensional theories lack, such as allowing for a representation of the quality of available information. It may thus be worth it to briefly discuss an increasingly popular and successful two-dimensional theory, one which makes use of lower and upper probabilities: imprecise probability (Coolen et al. 2010).

The term 'imprecise probability' is used to refer to a host of models in the literature. The commonality between these models is the use of some lower and upper probability, as opposed to a precise single value. The focus here will be on the theory developed by Peter Walley, who coined the name 'imprecise probability', and who was one of the first to provide the theory with mathematical and philosophical foundations (Walley 2000), along with Kuznetsov (1991) and Weichselberger (2000). Throughout this appendix, the term 'imprecise probability' is used to refer to this model specifically.

The reason for this focus on Walley's model is twofold. Firstly, it is mathematically general, and includes many other imprecise probability models as special cases. The Dempster-Shafer belief-functions (Shafer 1976) and Dubois possibility theory (Dubois and Prade 1988) mentioned in section 4.1.1 are examples of such models. Secondly, there is a clear interpretation of Walley's theory in terms of betting transactions, which has close ties to de Finetti's justification for subjective Bayesian probability, mentioned in section 2.3 (de Finetti 1931; de Finetti 1937). The purpose of this appendix is merely to be a brief

note on Walley's model. Miranda (2008) provides a more comprehensive overview of the mathematics, applications, and connection to de Finetti's work.

The basic idea of Walley's imprecise probability theory is that one's belief in a proposition A should be represented by both a lower probability $\underline{P}(A)$ and and upper probability $\overline{P}(A)$. Informally, one may interpret $\underline{P}(A)$ as the evidence definitely in favour of A, and $1-\overline{P}(A)$ as the evidence certainly against A (or equivalently in favour of the complement A^C) (Coolen et al. 2010). In terms of de Finetti's work, these lower and upper probabilities may be interpreted as special cases of lower and upper previsions. That is, the lower probability marks the cost/returns-ratio at which a rational decision maker should definitely buy a bet, while the upper probability is the ratio at which the bet should be sold. A fair price is one where the decision maker is willing to either sell or buy the bet, which means the ratios are equal. This forces $\underline{P}(A) = \overline{P}(A)$, such that the existence of a fair price leads to precise probabilities.

The special case $\underline{P}(A) = \overline{P}(A)$ represents those situations in which a precise probability for A may be determined, and thus Walley's model contains standard probability theory as a special case. The other extreme with $\underline{P}(A) = 0$ and $\overline{P}(A) = 1$ portrays complete ignorance. Narrow probability intervals are only justified by sufficient available information, while wide intervals represent more ignorance. The use of probability intervals thus admits a more general description: intervals allow for the same representations of belief as one-dimensional probability theory does, but additionally display the quality of the information at hand (Miranda 2008).

The case of total ignorance presents a illustrative example. In one-dimensional objective Bayesian probability theory one would use the principle of indifference to assign $P(A) = 0.5$ when no information is available. Contrast imprecise probability theory, which assigns $\underline{P}(A) = 0$ and $\overline{P}(A) = 1$. It is clear that the second contains more information, as it distinguishes this situation from that with $\underline{P}(A) = \overline{P}(A) = 0.5$: i.e. when there is definite information that allows for a reasonable belief that A is true with probability 0.5. In other words it may be true that, following some objective technique, the unique proper measure of reasonable belief in such cases is $P(A) = 0.5$ according to the one-dimensional objective Bayesian view, but such an assignment necessarily lacks information that distinguishes this case from those where information is available that definitely suggests $P(A) = 0.5$. One-dimensional theories do not contain information about the quality of evidence that leads to a probability assignment, which is important information to possess when one plans on acting on such beliefs.

It thus seems that imprecise probability theory has definite advantages over one-dimensional theories, especially in the context of decision theory. The theory is, however, not without downsides. Its lack of universal comparability may be a disadvantage when comparing the likelihood of different propositions, since it is not clear that a unique 'most likely' value $P(A)$ can always be determined from $\underline{P}(A)$ and $\overline{P}(A)$. Furthermore, the connection between information and imprecise probability is not yet fully apparent, and many frequently used statistical methods have not yet been generalised to employ two-dimensional probabilities (Coolen et al. 2010).

These are likely the most relevant advantages and disadvantages for our purposes, but the above is far from an exhaustive list. For a more complete overview of imprecise probability theory's strengths and challenges we refer to Miranda (2008, section 7).

B.3 The Ellsberg paradox

There exists a famous paradox in decision theory, called the Ellsberg paradox, which may pose a problem for Cox's result that Bayesian probability theory is the correct method for reasoning under uncertainty. The paradox involves a decision problem on which the decision by most human subjects can be shown to be inconsistent with our current formalism of Bayesian reasoning. This in itself should not be problematic, since - as Jaynes already noted in section 4.1.1 - *"human brains do many absurd things while failing to do many sensible things"*.

However, many experts in decision theory do not necessarily agree that a failure of the human mind is to blame here. Even Leonard Savage, who developed the theory of subjective expected utility (which incorporates Bayesian probability) that is directly violated by the paradox, continued to make the decision that violated his own postulates, even after this was pointed out to him (Ellsberg 1961).

That said, there exist a variety of possible resolutions to the Ellsberg paradox that do not violate Bayesian probability. Before we discuss these, let us examine the paradox in more detail.

Imagine an urn containing 90 balls. Of these 30 are red, and the remaining 60 are either blue or green in unknown proportion. We are requested to draw a ball from the urn at random, and bet on the colour of this ball. In particular, the choice is between the following two bets:

A) Receive \$100 if you draw a red ball.
B) Receive \$100 if you draw a blue ball.

After making a choice - without drawing a ball - we are offered a second bet from the same urn as follows:

C) Receive \$100 if you draw a red or green ball.
D) Receive \$100 if you draw a blue or green ball.

The common response is to select bet A in the first situation, and bet D in the second. These choices are inconsistent with Bayesian probability assignments, leading to the paradox. More specifically, the choices are inconsistent with subjective expected utility, which combines utility with Bayesian probability assignments. For these purposes it suffices to know that utility simply refers to the concept that every agent has a utility function U that assigns some value to a situation (e.g. winning \$100), assigning higher values to outcomes that are preferred more by the agent. Subjective expected utility in the simplest terms states that humans act to maximise their expected utility; i.e. maximise the value of U in expectation. The utility function can in principle take any form, though in our example it is often the case that for instance $U(\$100) > U(\$0)$, which simply states that one would rather have \$100 than nothing, ceteris paribus.

To understand the paradox, note that choosing bet A implies that one's belief state considers $P(\text{red})$ (the probability that the drawn ball is red) to be larger than $P(\text{blue})$. This further implies that one's choice in the second case should be to take bet C, since it must follow that $P(\text{red or green}) > P(\text{blue or green})$. This reasoning still works if for some reason one would rather lose the bet; i.e. one's utility function U considers $U(\$100)$ to be less than $U(\$0)$.

This last fact is perhaps better illustrated with a mathematical demonstration. We denote by R, B, and G respectively the probabilities that the drawn ball is red, blue, or green.

Then for bet A:

$$\begin{aligned} R \cdot U(\$100) + (1-R) \cdot U(\$0) &> B \cdot U(\$100) + (1-B) \cdot U(\$0) \\ R \cdot [U(\$100) - U(\$0)] &> B \cdot [U(\$100) - U(\$0)] \,. \end{aligned} \tag{B.7}$$

Assuming $U(\$100) > U(\$0)$ this simplifies to $R > B$, and if instead $U(\$100) < U(\$0)$ we have $B > R$. Then for bet D we obtain (note that all terms containing G cancel):

$$\begin{aligned} B \cdot U(\$100) + (1-B) \cdot U(\$0) + R \cdot U(\$0) + (1-R) \cdot U(\$100) > \\ B \cdot U(\$0) + (1-B) \cdot U(\$100) + R \cdot U(\$100) + (1-R) \cdot U(\$0). \end{aligned} \tag{B.8}$$

Which simplifies to:

$$B \cdot [U(\$100) - U(\$0)] > R \cdot [U(\$100) - U(\$0)] \,. \tag{B.9}$$

Note that this directly contradicts equation (B.7), regardless of the specific utility function U. This tendency to seek out situations in which probabilities are clearly defined has been termed ambiguity aversion. Ambiguity is characterised by unknown probability distributions, such as those describing the blue and green balls above. It is not to be confused with risk aversion, which is the tendency to pursue options that promise more certain payoff over uncertain payoff with higher expected value. Someone who is risk averse may in the above example select a guaranteed payoff of strictly less than \$10 over a payoff of \$30 for a drawn red ball.

The Ellsberg paradox shows that human behaviour is subject to ambiguity aversion to the extent that it can violate subjective expected utility and by extension Bayesian probability. Note that risk aversion does not violate these premises, since it may be explained by utility functions that grow more slowly with increasing payoff.

The obvious resolution to Ellsberg's paradox is that the preference of bet A over B together with bet D over C is inconsistent, and that some human failure is at work. Indeed, from the perspective that Bayesian probability is the correct method for reasoning under uncertainty, this should surely be the conclusion. As noted however, some experts - Savage himself among them - do not agree that such ambiguity aversion is irrational.

In fact, using the theory of imprecise probability discussed above in section B.2 it is possible to explain this behaviour in a consistent manner (Weichselberger 2000; Weichselberger and Augustin 1998). Intuitively this may be understood by noting that any probability assignment for the blue and green balls will be very uncertain (i.e. ambiguous), a situation that imprecise probability theory was precisely developed to tackle. We refer to Weichselberger and Augustin (1998) for a more comprehensive discussion than this.

Contrarily, there also exist a number of resolutions for the paradox that do not violate Bayesian probability. Indeed, in the strictest sense the above may not be a paradox from an objective Bayesian perspective. Invoking the principle of indifference suggests that $R = B = G$ in the above. Furthermore, a choice of bet A over B may suggest both a preference ($R > B$) or indifference ($R = B$). Under the latter assumption equations (B.7) and (B.9) should be extended to include equality, such that they are no longer necessarily contradictory. The preference of A and C - while still suggesting ambiguity aversion

- does therefore not violate Bayesian probability in Ellsberg's example. An interesting experiment may be to take another survey with the change that there are now 61 blue and green balls total. If the observed preferences do not change then that would indicate a more definite violation of Bayesian probability.

A second possible explanation is that this type of experiment triggers a deceit aversion mechanism. In many real-life situations not being told the probability of an event may serve as a deception to the person choosing between such events. It is conceivable that this intuitive distrust is projected onto the experimenter, such that the 30/90 probability of the red balls is compared to the lower end of the 0/90 to 60/90 range of the blue balls. Similarly, the 60/90 range of the blue and green balls may be compared to the lower end of the 30/90 to 90/90 range of the red and green balls. Despite the fact that the same urn is used in both experiments, it is quite possible that the experimental subjects do not intuitively realise that the urn cannot be modified between draws, or that they would be afraid to be deceived on this front in real-life situations as well.

The Ellsberg paradox provides an interesting insight into human behaviour, but while its results seems to violate subjective expected utility and perhaps Bayesian probability, it is not clear that these violations are not simply motivated by irrational behaviour. It is however noteworthy that this possible irrational behaviour vanishes under application of imprecise probability theory, which may constitute another advantage it enjoys over one-dimensional Bayesian probability.

B.4 Bayes's theorem and objective Bayesian principles

The principal theorem in Bayesian probability - and its namesake - is that of Bayes. Bayes's theorem governs in what precise manner one should update their probabilities in light of new data and evidence. This updating is one of the cornerstones of objective Bayesian philosophy, as it provides a way to uniquely determine how one's reasonable belief should be influenced by new information.

However, in his 1979 paper Teddy Seidenfeld argues that Bayes's theorem is incompatible with the principle of maximum entropy and transformation groups. If accurate, this claim certainly poses a problem for the objective Bayesian perspective - particularly Jaynes's perspective - as these formalisms represent two chief methods for determining objective priors. Let us therefore consider the argument by Seidenfeld in more detail.

As mentioned in Appendix A.3, the principle of transformation groups is fairly reminiscent of Jeffreys's use of invariants. Both approaches consider the statistical model in question, and base their prior distributions on some family of transformations that arises from the model's symmetries. Thus, Seidenfeld suggests, different experiments - that use different models - performed to estimate the same parameter may lead one to assign dissimilar total ignorance (i.e. indifference) priors (Seidenfeld 1979).

Consider the case where two separate experiments possessing differing symmetries - A and B - are available. By Bayes's theorem the order in which these experiments are performed should not matter for the final posterior distribution gathered from the prior and the data. However, doing experiment A first will lead one to use an invariant prior suitable to the symmetries in A, while contrarily doing experiment B first may leave one with a different form of invariant prior suitable to B. This prior is then in both cases updated with the same data gathered by experiments A and B, but if the above implementations started with distinct priors, then they will not find the same posterior.

Bayes's theorem guarantees that the order in which data is gathered is irrelevant for any posterior, and yet this order matters for the total ignorance prior obtained by invariance. We present a more concrete example of this that was provided by Seidenfeld in the following.

Suppose that the goal of an experimenter is to investigate the volume v of some hollow cube and that there are two experiments they may perform to do so. First, they may fill the cube with some liquid of known density, and weigh this same amount of liquid on a scale with some (perhaps known) variance. Since the volume of the cube proportional to the magnitude of the volume v, invariance requires adoption of a informationless prior that is uniform over possible values of v.

Second, the experimenter may take a rod of known density and cut it to the length of the cube's edge, before weighing the segment on the scale. Here invariance dictates that the uniform prior be over the possible weights of the rod, and thus the total ignorance prior will be uniform over the lengths of the cube's edge. However, the volume v of a cube is related to its edge length l by the relation $v = l^3$, and so they have produced a prior uniform over $\sqrt[3]{v}$.

Thus, depending on which experiment was analysed first a different total ignorance prior will be found, and after tallying up the available evidence there will be two distinct posterior distributions for v. Such a result is not acceptable for objective Bayesian reasoning, as it is built on the principle that there exists a single - unique - representation of reasonable belief. Seidenfeld suggests that the 'fatal flaw' is the wish to represent ignorance by a single precise probability. Indeed, this argument provides a strong case for the superiority of two-dimensional theories of probability, such as the imprecise probability theory of section B.2. In such a theory one represents ignorance by a wide probability interval, as opposed to by a single value, and thus avoids the issues presented here.

Certainly, the inability to provide completely defensible representations of ignorance has been an issue with the one-dimensional objective Bayesian philosophy. Fortunately however, it is rarely the case that one is completely ignorant of all knowledge relevant to the problem at hand. In such cases, methods for determining objective priors that represent the known information are available, a number of which are discussed in Appendix A. When this information is expressed as an expected value of some functions of the unknown variables, the principle of maximum entropy - discussed in Appendix A.2 - may be utilised.

According to Seidenfeld however, MaxEnt is no more consistent with Bayes's theorem than the aforementioned principles of invariance (which encompass both Jeffreys's methods, as well as Jaynes's transformation groups) are. He claims the maximum entropy principle can output distributions that contain more information than is actually available, and supports this assertion by means of a fictional conversation between two statisticians; J. (a Bayesian who shares the views of Jaynes) and B. (a non-objective Bayesian).

Specifically, Seidenfeld provides an example in which the maximum entropy principle produces too informed distributions in the presence of nuisance parameters. Since example itself is rather involved, we refer to his paper for a full treatment (Seidenfeld 1979). Here we present solely the essence of the argument, which is as follows: if the experimenter's initial belief state considers the parameter of interest and any nuisance factors independently, then directly maximising entropy subject to some constraints may not correctly take the uncertainty in these nuisance factors into account. In particular, the process may neglect the uncertainty that arises from the interaction between the uncertainty in the nuisance factor and uncertainty in the parameter of interest.

In Seidenfeld's account J. was supplied with some constraints corresponding to a posterior distribution of B.'s, such that he could estimate that posterior using MaxEnt. After learning B.'s experimental process - specifically the number of independently performed measurements - J. was able to fix the nuisance parameter to a specific value; something B. was unable to do by updating on his prior with Bayes's theorem, even though B. possessed still more information than J. Moreover, the value J. fixed for the nuisance factor was not the actual value B. determined later; an error that resulted from maximum entropy neglecting the interdependence of this factor with the parameter of interest.

It thus seems that MaxEnt cannot accurately handle situations in which multiple parameters are present whose uncertainties may interact to create more uncertainty, especially if the initial belief state does not already specify these interactions. That said, there are authors who suggest that conditionalising on evidence via Bayes's theorem should not be a central principle of objective Bayesian methods, instead relying on MaxEnt to continually update one's reasonable belief. One such author, Jon Williamson, writes (Williamson 2011):

> *"This is 'Bayesianism' in the sense that probabilities are are construed as degrees of belief, as in Bayes (1764), not in the sense that probabilities are updated by Bayesian conditionalisation: as will be explained in §3, the objective Bayesian does not need to advocate Bayesian conditionalisation, and indeed the version of objective Bayesianism presented here explicitly rejects conditionalisation."*

This statement may be surprising, considering that the 'conditionalisation' Willamson mentions is given by Bayes's theorem, which itself is essentially the condition that the product rule of probability theory should be consistent (to see this, note that $P(A|B,I)\,P(B|I) = P(A,B|I) = P(B|A,I)\,P(A|I)$ where A, B are some propositions and I is a state of information is a sufficient condition to derive Bayes's theorem). The arguments provided in Williamson's work are however beyond both the scope of this thesis and the understanding of this author. We thus refer to his work and encourage any interested parties to evaluate it for themselves.

In this section some additional arguments against both objective Bayesian invariance principles and MaxEnt were discussed. Particularly, it seems that situations can be found where these principles violate the conditionalisation given by Bayes's theorem. Whether this implies a fault in objective Bayesian methods alone, or some more fundamental flaw in the philosophy of representing belief by a single number is not clear. Indeed, it is not yet definitive that the error lies in the objective Bayesian approach at all; conditionalisation may instead be unwarranted, although this may seem implausible given its prominent position in Bayesian methods.

Bibliography

[1] C. Abellán et al. "Generation of Fresh and Pure Random Numbers for Loophole-Free Bell Tests". In: *Physical Review Letters* 115.25 (2015), pp. 250–403. DOI: http://dx.doi.org/10.1103/PhysRevLett.115.250403.

[2] J. Aczél. *Lectures on Functional Equations and their Applications*. Academic Press, 1966. ISBN: 978-0080955254.

[3] J. Aldrich. "R. A. Fisher on Bayes and Bayes Theorem". In: *Bayesian Analysis* 3.1 (2008), pp. 161–170.

[4] S. Arnborg and G. Sjödin. *A Note on Foundations of Bayesianism*. Tech. rep. Royal Institute of Technology, Stockholm, 1999.

[5] A. Aspect, P. Grangier, and G. Roger. "Experimental Tests of Realistic Local Theories via Bell's Theorem". In: *Physical Review Letters* 47.7 (1981), pp. 460–463. DOI: 10.1103/PhysRevLett.47.460.

[6] A. Aspect, P. Grangier, and G. Roger. "Experimental Realization of Einstein-Podolsky-Rosen-Bohm Gedankenexperiment: A New Violation of Bell's Inequalities". In: *Physical Review Letters* 49.2 (1982), pp. 91–94. DOI: 10.1103/PhysRevLett.49.91.

[7] A. Aspect, J. Dalibard, and G. Roger. "Experimental Test of Bell's Inequalities Using Time-Varying Analyzers". In: *Physical Review Letters* 49.25 (1982), pp. 1804–1807. DOI: 10.1103/PhysRevLett.49.1804.

[8] L. E. Ballentine. "The Statistical Interpretation of Quantum Mechanics". In: *Reviews of Modern Physics* 42.4 (1970). DOI: 10.1103/RevModPhys.42.358.

[9] J. S. Bell. "On the Einstein Podolsky Rosen Paradox". In: *Physics* 1 (1964), pp. 195–200.

[10] J. O. Berger and J. M. Bernardo. *On the Development of the Reference Prior Method*. Tech. rep. Department of Statistics, Purdue University, 1991.

[11] J. M. Bernardo. "Reference Posterior Distributions for Bayesian Inference". In: *Journal of the Royal Statistical Society* 41.2 (1979), pp. 113–147. DOI: 10.1103/PhysRev.106.620.

[12] J. M. Bernardo. *A Bayesian Mathematical Statistics Primer*. Tech. rep. 2006. eprint: http://iase-web.org/documents/papers/icots7/3I2_BERN.pdf.

[13] C. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006. ISBN: 978-0387310732.

[14] D. Bohm. "A Suggested Interpretation of the Quantum Theory in Terms of "Hidden" Variables. I". In: *Physical Review* 85.2 (1952), pp. 166–179.

[15] D. Bohm and Y. Aharonov. "Discussion of Experimental Proof for the Paradox of Einstein, Rosen, and Podolsky". In: *Physical Review* 108.4 (1957), p. 1070.

[16] L. Boltzmann. "Analytischer Beweis des zweiten Hauptsatzes der mechanischen Wärmetheorie aus den Stzen über das Gleichgewicht der lebendigen Kraft". In: *Wiener Berichte* 63 (1 1882), pp. 712–732.

[17] L. Boltzmann. "Einige allgemeine Sätze über Wärmegleichgewicht". In: *Wiener Berichte* 63 (1 1882), pp. 679–711.

[18] L. Boltzmann. "Über das Wärmegleichgewicht zwischen mehratomigen Gasmoleküle". In: *Wiener Berichte* 63 (1 1882), pp. 397–418.

[19] S. Boucheron and E. Gassiat. "A Bernstein-Von Mises Theorem for discrete probability distributions". In: *Electronic Journal of Statistics* 3 (2009), pp. 114–148. DOI: 10.1214/08-EJS262.

[20] G. E. P. Box and N. R. Draper. *Empirical Model-Building and Response Surfaces*. Wiley, 1987. ISBN: 978-0471810339.

[21] K. P. Burnham and D. R. Anderson. *Model Selection and Multi-Model Inference*. Springer, 2002. ISBN: 978-0387953649.

[22] J. F. Clauser et al. "Proposed Experiment to Test Local Hidden-Variable Theories". In: *Physical Review Letters* 23 (1969), p. 880. DOI: 10.1103/PhysRevLett.23.880.

[23] R. Colbeck and R. Renner. "No extension of quantum theory can have improved predictive power". In: *Nature Communications* 2 (2011). DOI: 10.1038/ncomms1416. arXiv: 1208.4123 [quant.ph].

[24] F. P. A. Coolen, M. C. M. Troffaes, and T. Augustin. "Imprecise Probabilities". In: *International Encyclopedia of Statistical Sciences* (2010).

[25] A. A. Cournot. *Exposition de la théorie des chances et des probabilités*. L. Hachette, Paris, 1843.

[26] D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman & Hall, 1974. ISBN: 978-0412161605.

[27] R. T. Cox. "Probability, frequency and reasonable expectation". In: *American Journal of Physics* 17 (1946), pp. 1–13. DOI: 10.1119/1.1990764.

[28] R. T. Cox. *The Algebra of Probable Inference*. Johns Hopkins University Press, 1961. ISBN: 9780801869822.

[29] R. T. Cox. "Of Inference and Inquiry: An Essay in Inductive Logic". In: *The Maximum Entropy Formalism* (1978).

[30] U. Daepp and P. Gorkin. *Reading, Writing, and Proving: A Closer Look at Mathematics*. Springer, 2003. ISBN: 978-1441994783.

[31] Merriam-Webster Dictionary. *Entry for 'epistemology'*. 1996. URL: http://www.merriam-webster.com/dictionary/epistemology (visited on 04/02/2016).

[32] Merriam-Webster Dictionary. *Entry for 'ontology'*. 1996. URL: http://www.merriam-webster.com/dictionary/ontology (visited on 04/02/2016).

[33] D. Dubois and H. Prade. *Possibility Theory: an Approach to Computerized Processing of Uncertainty*. Plenum Press, 1988. ISBN: 0306425203.

[34] R. E. Ellis. "On the Foundations of the Theory of Probabilities". In: *Transactions of the Cambridge Philosophical Society* 8 (1843).

[35] R. E. Ellis. "Remarks on the Fundamental Principles of the Theory of Probabilities". In: *Transactions of the Cambridge Philosophical Society* 9 (1854).

[36] D. Ellsberg. "Risk, Ambiguity, and the Savage Axioms". In: *The Quarterly Journal of Economics* 75.4 (1961), pp. 643–669.

[37] H. Everett. “Theory of the Universal Wavefunction”. PhD thesis. Princeton University, 1956.

[38] H. Everett. “Relative State Formulation of Quantum Mechanics”. In: *Reviews of Modern Physics* 29 (1957), pp. 454–462. DOI: 10.1103/RevModPhys.29.454.

[39] S. E. Fienberg. “A Brief History of Statistics in Three and a One-Half Chapters: A Review Essay”. In: *Statistical Science* 7.2 (1992), pp. 208–225.

[40] S. E. Fienberg. “When Did Bayesian Inference Become “Bayesian”?” In: *Bayesian Analysis* 1.1 (2006), pp. 1–40.

[41] T. L. Fine. *Theories of Probability: An Examination of Foundations*. Academic Press, 1973. ISBN: 978-1483238791.

[42] B. de Finetti. “Sul signifcato soggettivo della probabilità”. In: *Fundamenta Mathematicae* 17 (1931), pp. 298–319.

[43] B. de Finetti. “La prévision: ses lois logiques, ses sources subjectives”. In: *Annales de l’Institut Henri Poincaré* 7 (1937), pp. 1–68.

[44] D. Freedman. “On the Bernstein-von Mises Theorem with Infinite Dimensional Parameters”. In: *The Annals of Statistics* 27.4 (1999), pp. 1119–1140.

[45] S. J. Freedman and J. F. Clauser. “Experimental Test of Local Hidden-Variable Theories”. In: *Physical Review Letters* 28.14 (1972), p. 938. DOI: http://dx.doi.org/10.1103/PhysRevLett.28.938.

[46] M. Ghosh. “Objective Priors: An Introduction for Frequentists”. In: *Statistical Science* 26.2 (2011), pp. 187–202.

[47] J. W. Gibbs. *Elementary Principles in Statiscial Mechanics*. Charles Scribner’s Sons, 1902.

[48] G. Gigerenzer et al. *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge University Press, 1989. ISBN: 978-0521398381.

[49] R. D. Gill. “Time, Finite Statistics, and Bells Fifth Position”. In: (2003). arXiv: 0301059 [quant.ph].

[50] J. Y. Halpern. “A counterexample to theorems of Cox and Fine”. In: *Proceedings, Association for the Advancement of Artificial Intelligence* 10 (1996), pp. 1313–1319.

[51] J. Y. Halpern. “Cox’s theorem revisited: technical addendum”. In: *Journal of Artificial Intelligence Research* 11 (1999), pp. 429–435.

[52] G. ’t Hooft. *Why do people categorically dismiss some simple quantum models?* 2012. URL: https://physics.stackexchange.com/questions/34217/why-do-people-categorically-dismiss-some-simple-quantum-models (visited on 02/15/2016).

[53] K. S. Van Horn. “Constructing a logic of plausible inference: a guide to Cox’s theorem”. In: *International Journal of Approximate Reasoning* 34 (2003), pp. 3–24. DOI: 10.1016/S0888-613X(03)00051-3.

[54] V. S. Huzurbazar. *Sufficient Statistics*. CRC Press, 1976. ISBN: 978-0824762964.

[55] E. T. Jaynes. “Information Theory and Statistical Mechanics I”. In: *The Physical Review* 106.4 (1957), pp. 620–630. DOI: 10.1103/PhysRev.106.620.

[56] E. T. Jaynes. “Information Theory and Statistical Mechanics II”. In: *The Physical Review* 108 (1957). DOI: 10.1103/PhysRev.108.171.

[57] E. T. Jaynes. “Information Theory and Statistical Mechanics”. In: *Statistical Physics*. Ed. by K.W. Ford. W. A. Benjamin, 1963, pp. 181–218.

[58] E. T. Jaynes. “Gibbs vs Boltzmann Entropies”. In: *American Journal of Physics* 33.5 (1965), pp. 391–398. DOI: 10.1119/1.1971557.

[59] E. T. Jaynes. “Foundations of Probability Theory and Statistical Mechanics”. In: *Studies in the Foundations Methodology and Philosophy of Science - Volume 1*. Ed. by M. Bunge. Springer-Verlag, 1967, pp. 77–101. DOI: 10.1007/978-3-642-86102-4_6.

[60] E. T. Jaynes. “Prior Probabilities”. In: *IEEE Transactions On Systems Science and Cybernetics* 4.3 (1968), pp. 227–241.

[61] E. T. Jaynes. “The Well-Posed Problem”. In: *Foundations of Physics* 3 (1973), pp. 477–493. DOI: 10.1007/BF00709116.

[62] E. T. Jaynes. “Where do we stand on Maximum Entropy?” In: *Proceedings, Maximum Entropy Formalism Conference* (1978). eprint: http://bayes.wustl.edu/etj/articles/stand.on.entropy.pdf.

[63] E. T. Jaynes. “Bayesian Methods, General Background”. In: *Proceedings, Maximum Entropy and Bayesian Methods in Applied Statistics* (1984).

[64] E. T. Jaynes. “The Evolution of Carnot’s Principle”. In: *Proceedings, EMBO Workshop on Maximum-Entropy Methods* (1984). eprint: http://bayes.wustl.edu/etj/articles/ccarnot.pdf.

[65] E. T. Jaynes. “Clearing Up Mysteries - The Original Goal”. In: *Proceedings, Maximum Entropy and Bayesian Methods* (1989).

[66] E. T. Jaynes. “Probability Theory as Logic”. In: *Proceedings, Maximum Entropy and Bayesian Methods* (1990).

[67] E. T. Jaynes. “Macroscopic Prediction”. In: *Complex Systems - Operational Approaches in Neurobiology, Physics, and Computers*. Ed. by H. Haken. Springer-Verlag, 1996, pp. 254–269.

[68] E. T. Jaynes. “Probability in Quantum Theory”. In: *Proceedings, Complexity, Entropy, and the Physics of Information* (1996).

[69] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003. ISBN: 978-0521592710.

[70] E. T. Jaynes and F. W. Cummings. “Comparison of quantum and semiclassical radiation theories with application to the beam maser”. In: *Proceedings, Institute of Electrical and Electronics Engineers* 51.1 (1963), pp. 89–109. DOI: 10.1109/PROC.1963.1664.

[71] H. Jeffreys. *Scientific Inference*. Cambridge University Press, 1931.

[72] H. Jeffreys. *The Theory of Probability*. Oxford University Press, 1939. ISBN: 978-0191589676.

[73] H. Jeffreys. “An Invariant Form for the Prior Probability in Estimation Problems”. In: *Proceedings of the Royal Society of London* 186.1007 (1946), pp. 453–461.

[74] H. Jeffreys. “The Present Position in Probability Theory”. In: *British Journal for Philosophy of Science* 5 (1955), pp. 275–289.

[75] M. Kac. “Some Stochastic Problems in Physics and Mathematics”. In: *Colloquium Lectures in Pure and Applied Science no.2*. Socony-Mobil Oil Company, 1956.

[76] R. E. Kass and L. Wasserman. “The Selection of Prior Distributions by Formal Rules”. In: *Journal of the American Statistical Association* 91.435 (1996). DOI: 10.1080/01621459.1996.10477003.

[77] M. G. Kendall and D. G. Stuart. *The Advanced Theory of Statistics. Vol 2: Inference and Relationship*. Griffin, London, 1973. ISBN: 978-0852642559.

[78] A. N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, 1933. ISBN: 978-3642498886. DOI: 10.1007/978-3-642-49888-6.

[79] D. Kuić, P. Županović, and D. Juretić. "Macroscopic Time Evolution and MaxEnt Inference for Closed Systems with Hamiltonian Dynamics". In: *Foundations of Physics* 42.2 (2012), pp. 319–339.

[80] V. P. Kuznetsov. "Interval Statistical Models". In: *Radio i Svyaz' Publishing House* (1991).

[81] J. Å. Larsson. "Loopholes in Bell inequality tests of local realism". In: *Journal of Physics A* 47 (2014).

[82] D. Malakoff. "Bayes Offers a 'New' Way to Make Sense of Numbers". In: *Science* 286 (1999), pp. 1460–1464. DOI: 10.1126/science.286.5444.1460.

[83] J. C. Maxwell. "On Boltzmann's Theorem on the Average Distribution of Energy in a System of Material Points". In: *Journal of Science* 14 (1882), pp. 299–312.

[84] S. B. McGrayne. *The Theory That Would Not Die*. Yale University Press, 2012. ISBN: 978-0300188226.

[85] E. Miranda. "A survey of the theory of coherent lower previsions". In: *International Journal of Approximate Reasoning* 48.2 (2008), pp. 628–658. DOI: 10.1016/j.ijar.2007.12.001.

[86] Nathaniel. *Ignorance in statistical mechanics*. Physics Stack Exchange (version: 2012-03-10). 2012. URL: http://physics.stackexchange.com/q/21960.

[87] R. O'Connor. *Bells Casino*. 2015. URL: http://r6.ca/blog/20151210T033709Z.html (visited on 02/15/2016).

[88] J. B. Paris. *The Uncertain Reasoner's Companion: a Mathematical Perspective*. Cambridge University Press, 1994. ISBN: 978-0521032728.

[89] S. L. van der Pas. "Much ado about the p-value". BS Thesis. Leiden University, 2010.

[90] J. C. Polkinghorne. *The Quantum World*. Princeton University Press, 1986. ISBN: 978-0691023885.

[91] R. Pool. "Chaos Theory: How Big an Advance?" In: *Science* 245 (1989), pp. 26–28.

[92] F. Porter. *Density Matrix Formalism, Course Notes*. 2004. eprint: http://www.cithep.caltech.edu/~fcp/physics/quantumMechanics/densityMatrix.

[93] C. P. Robert, N. Chopin, and J. Rousseau. "Harold Jeffreyss Theory of Probability Revisited". In: *Statistical Science* 24.2 (2009), pp. 141–172. DOI: 10.1214/09-STS284.

[94] L. J. Savage. *The Foundations of Statistics*. Dover Publications, 1954. ISBN: 978-0486623498.

[95] T. Seidenfeld. "Why I am not an objective Bayesian; some reflections prompted by Rosenkrantz". In: *Theory and Decision* 11.4 (1979), pp. 413–440.

[96] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976. ISBN: 978-0691100425.

[97] C. E. Shannon. "A Mathematical Theory of Communication". In: *Bell System Technical Journal* 27.379 (1948), p. 623.

[98] A. Shimony, M. A. Horne, and J. F. Clauser. "Comment on the theory of local beables". In: *Epistemological Letters* 13.1 (1976).

[99] P. Smets. "The Transferable Belief Model and other Interpretations of Dempster-Shafer's Model". In: *Proceedings, Sixth Conference on Uncertainty in Artificial Intelligence* (1990). arXiv: 1304.1120v1 [cs.AI].

[100] P. Snow. "On the correctness and reasonableness of Cox's theorem for finite domains". In: *Computational Intelligence* 14.3 (1998), pp. 452–459.

[101] P. Snow. "The Disappearance of Equation Fifteen, a Richard Cox Mystery". In: *Proceedings, FLAIRS Conference* (2002), pp. 602–606.

[102] S. M. Stigler. *The History of Statistics*. Harvard University Press, 1986.

[103] M. Stone. "The theory of representations of Boolean algebras". In: *Transactions of the American Mathematical Society* 40 (1936), pp. 37–111.

[104] A. Terenin and D. Draper. "Rigorizing and Extending the Cox-Jaynes Derivation of Probability: Implications for Statistical Practice". In: (2015). arXiv: 1507.06597v1 [math.ST].

[105] M. Tribus. *Rational Descriptions, Decisions and Designs*. Pergamon Press, 1969. ISBN: 978-0080063935.

[106] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000. ISBN: 978-0521784504.

[107] S. Vaughan. *Scientific Inference: Learning from Data*. Cambridge University Press, 2013. ISBN: 978-1107607590.

[108] A. Wald. "Statistical Decision Functions". In: *Annals of Mathematical Statistics* 20.2 (1949), pp. 165–205.

[109] P. Walley. "Towards a unified theory of imprecise probability". In: *International Journal of Approximate Reasoning* 24.2-3 (2000), pp. 125–148. DOI: 10.1016/S0888-613X(00)00031-1.

[110] K. Weichselberger. "The theory of interval-probability as a unifying concept for uncertainty". In: *International Journal of Approximate Reasoning* 24.2-3 (2000), pp. 149–170. DOI: 10.1016/S0888-613X(00)00032-3.

[111] K. Weichselberger and T. Augustin. "Analysing Ellsberg's paradox by means of interval-probability". In: *Econometrics in Theory and Practice*. Ed. by R. Galata and H. Küchenhoff. Physika-Verlag, 1998, pp. 291–304.

[112] E. P. Wigner. "Reminiscences on Quantum Theory". Colloquium talk at Washington University. 1974.

[113] Wiktionary. *Entry for 'epistemic'*. 2016. URL: https://en.wiktionary.org/wiki/epistemic (visited on 10/04/2016).

[114] J. Williamson. "Objective Bayesianism, Bayesian conditionalisation and voluntarism". In: *Synthese* 178.1 (2011), pp. 67–85.

[115] H. Wimmel. *Quantum Physics & Observed Reality: A Critical Interpretation of Quantum Mechanics*. World Scientific Pub Co Inc, 1992. ISBN: 978-9810210106.

[116] R. L. Wolpert. "A conversation with James O. Berger". In: *Statistical Science* 9 (2004), pp. 205–218.