Information Theory and Partial Belief Reasoning

by

Stefan Hermann Lukits

Master of Arts (M. A.), The University of British Columbia
Master of Science (Mag. rer. nat.), Karl-Franzens Universität Graz
(University of Graz)

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Philosophy)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

March 2016

Abstract

The dissertation investigates the nature of partial beliefs and norms governing their use. One widely accepted (though not uncontested) norm for partial belief change is Bayesian conditionalization. Information theory provides a far-reaching generalization of Bayesian conditionalization and gives it a foundation in an intuition that pays attention principally to information contained in probability distributions and information gained with new evidence.

This generalization has fallen out of favour with contemporary epistemologists. They prefer an eclectic approach which sometimes conflicts with norms based on information theory, particularly the entropy principles of information theory. The principle of maximum entropy mandates a rational agent to hold minimally informative partial beliefs given certain background constraints; the principle of minimum cross-entropy mandates a rational agent to update partial beliefs at minimal information gain consistent with the new evidence. The dissertation shows that information theory generalizes Bayesian norms and does not conflict with them.

It also shows that the norms of information theory can only be defended when the agent entertains sharp credences. Many contemporary Bayesians permit indeterminate credal states for rational agents, which is incompatible with the norms of information theory. The dissertation then defends two claims: (1) the partial beliefs that a rational agent holds are formally expressed by sharp credences; and (2) when a rational agent updates these partial beliefs in the light of new evidence, the norms used are based on and in agreement with information theory.

In the dissertation, I defuse a collection of counter-examples that have been marshaled against entropy principles. More importantly, building on previous work by others and expanding it, I provide a coherent and comprehensive theory of the use

of information theory in formal epistemology. Information theory rivals probability theory in formal virtue, theoretical substance, and coherence across intuitions and case studies. My dissertation demonstrates its significance in explaining the doxastic states of a rational agent and in providing the right kind of normativity for them.

Preface

This dissertation is original, unpublished, independent work by the author, with the exception of chapter 5 (published in Lukits, 2015) and chapter 7 (published in Lukits, 2014b).

Table of Contents

Appendices

List of Figures

Chapter 1

Introduction

1.1 Information Theory and Probability Kinematics

This dissertation defends two claims: (1) the partial beliefs that a rational agent entertains are formally expressed by sharp credences; and (2) when a rational agent updates these partial beliefs in the light of new evidence, the norms used are based on and in agreement with information theory.

A Bayesian framework for partial beliefs is assumed throughout. I will not explicitly argue for this framework, although the overall integrity of my claims hopefully serves to support the Bayesian approach to epistemology, especially the rules governing belief change. Some say that the Bayesian approach conflicts with information theory. I will show that there is no such conflict. The account that I defend is a specific version of Bayesian epistemology. The major task of this dissertation is therefore to convince a Bayesian that a consistent application of the principles and intuitions leading to the Bayesian approach will lead to the acceptance of norms which are based on or in agreement with information theory.

If an agent's belief state is representable by a probability distribution (or density), then Bayesians advocate formal and normative guidelines for changing it in the light of new evidence. Richard Jeffrey calls the investigation of these guidelines 'probability kinematics' (see Jeffrey, 1965). Standard conditioning is the updating procedure considered by Bayesians to be objective and generally valid. The $|$ operator (probabilistic conditioning) provides a unique posterior probability distribution which accords with both basic probability axioms and intuitions about desired properties of updated probabilities.

Although almost all the claims, examples, and formal methods presented in the following pertain to discrete and finite probability distributions, there are usually equivalent things to be said for distributions over infinite event spaces and for continuous probability densities. Sometimes this requires some mathematical manoeuvering (as is the case for the Shannon entropy), sometimes it requires no more than substitution of an integral sign for a sum sign. I will exclusively refer to finite probability distributions, but many of the claims can be extended to suit the continuous or infinite case.

Bayesian probability kinematics requires a prior probability distribution to precede any meaningful evaluation of evidence and considers standard conditioning to be mandatory for a rational agent when she forms her posterior probabilities, just in case her evidence is expressible in a form which makes standard conditioning an option. There are various situations, such as the *Judy Benjamin* problem or the *Brandeis Dice* problem (see example 8), in which standard conditioning does not appear to be an option (although some make the case that it is, so this point needs to be established independently). Therefore the question arises whether there is room for a more general updating method, whose justification will entail a justification of standard conditioning, but not be entailed by it.

E.T. Jaynes has suggested a unified updating procedure which generalizes standard conditioning called the principle of maximum entropy, or PME for short. PME additionally to Bayesian probability kinematics uses information theory to develop updating methods which keep the entropy of probability distributions high (in the synchronic case) and their cross-entropy low (in the diachronic case).

PME is based on a subjective interpretation of probabilities, where probabilities represent the degree of uncertainty, or the lack of information, of the agent holding these probabilities. In this interpretation, probabilities do not represent frequencies or objective probabilities, although there are models of how subjective probabilities relate to them. PME enjoys a wide range of applications in science (for summaries see Shore and Johnson, 1980, 26; Buck and Macaulay, 1991; Karmeshu, 2003; Debbah and Müller, 2005, 1668; and Klir, 2006, 95). Most Bayesians reject the notion that PME enjoys the same level of justification as standard conditioning and maintain that

there are cases in which it delivers results which a rational agent should not, or is not required to, accept.

Note that we distinguish between separate problems: on the one hand, there may be objective methods of determining probabilities prior to any evidence or observation (call these 'absolutely' prior probabilities), for example from some type of principle of indifference; on the other hand, there may be objective methods of determining posterior probabilities from given prior probabilities (which themselves could be posterior probabilities in a previous instance of updating, call these 'relatively' prior probabilities) in case standard conditioning does not apply.

Another way to distinguish between absolutely prior distributions and relatively prior distributions is to use Arnold Zellner's 'antedata' and 'postdata' nomenclature (see Zellner, 1988). My work is concerned with a logic of belief change, not with objectivism or convergence, although PME has been used to defend objectivism, notably by Jaynes himself (more about this in section 2.1).

Formal epistemologists widely concur that PME is beset with too many conceptual problems and counterexamples to yield a generally valid objective updating procedure. Against the tide of this agreement, I am mounting a defence of PME as a viable candidate for a generally valid, objective updating procedure. This defence also identifies problems with PME, especially in chapter 4, but I consider these problems to be surmountable. PME is unmatched by other updating procedures in its wide applicability, integration with other disciplines, and formal power (see chapter 4 for a comparison with the geometry of reason). Giving up on it means giving up on generality overall and settling with piecemeal approaches and a patchwork of ad hoc solutions to cases.

Many of the portrayals of PME's failings are flawed and motivated by a desire to demonstrate that the labour of the epistemologist in interpreting probability kinematics on a case-by-case basis is indispensable. This 'full employment theorem' of probability kinematicsis is widely promulgated by textbooks and passed on to students as expert consensus. There is a formally proven equivalent for the full employment theorem in computer science:

> **Full Employment Theorem** No algorithm can optimally perform a particular task done by some class of professionals.

The theorem ensures that a variety of tasks in computer science is performed by programs whose optimization is in principle an open question. There cannot be a proof that a particular program executes a task optimally. The theorem secures employment for computer programmers, for example in developing search engines, spam filters, or virus detection. Rudolf Carnap discusses the "impossibility of an automatic inductive procedure" (see Carnap, 1962, 192–199) making explicit reference to Alan Turing, who is behind the full employment theorem in computer science.

Teddy Seidenfeld makes the case that the appeal of PME is for particular cases with a view to pragmatic advantages, but that the general theory needs to recognize the problems and maintain a watchful eye, given specific cases, which formalism produces the best results:

> A pragmatic appeal to successful applications of the MAXENT formalism cannot be dismissed lightly. The objections that I raise in this paper are general. Whether (and if so, how) the researchers who apply MAXENT avoid these difficulties remains an open question. Perhaps, by appealing to extra, discipline-specific assumptions they find ways to resolve conflicts within MAXENT theory. A case-by-case examination is called for. (Seidenfeld, 1986, 262.)

A similar sentiment is expressed in Grove and Halpern, 1997, 210.

> Against this, a rule like cross-entropy seems extremely attractive. It provides a single general recipe which can be mechanically applied to a huge space of up dates [...] since all we do in this paper is analyze one particular problem, we must be careful in making general state ments on the basis of our results. Nevertheless, they do seem to support the claim that sometimes, the only 'right' way to update, especially in an incompletely specified situation, is to think very carefully about the real nature and origin of the information we receive, and then (try to) do whatever is necessary to find a suitable larger space in which we can condition. If this doesn't lead to conclusive

> results, perhaps this is because we do not understand the information well enough to update with it. However much we might wish for one, a generally satisfactory mechanical rule such as cross-entropy, which saves us all this questioning and work, probably does not exist.

This is echoed both in Halpern's textbook *Reasoning About Uncertainty*, where Halpern emphasizes that "there is no escaping the need to understand the details of the application" (2003, 423), and by Richard Bradley, who determines that much "will depend on the judgmental skills of the agent, typically acquired not in the inductive logic class but by subject specific training" (Bradley, 2005, 349).

Grove and Halpern, in their article about the *Judy Benjamin* problem, admonish the objectivists in accordance with the full employment theorem that "one must always think carefully about precisely what the information means" (Grove and Halpern, 1997, 6), and "the only right way to update, especially in an incompletely specified situation, is to think very carefully about the real nature and origin of the information we receive" (Grove and Halpern, 1997, 3). Igor Douven and Jan-Willem Romeijn advocate an innovative approach to the *Judy Benjamin* problem, in which epistemic entrenchment supplies a subjectivist perspective on these types of problem and delivers a Jeffrey partition so that instead of MAXENT Jeffrey's rule can be applied. They agree with Bradley that "even Bayes' rule 'should not be thought of as a universal and mechanical rule of updating, but as a technique to be applied in the right circumstances, as a tool in what Jeffrey terms *the art of judgment*.' In the same way, determining and adapting the weights [epistemic entrenchment] supposes, or deciding when Adams conditioning applies, may be an art, or a skill, rather than a matter of calculation or derivation from more fundamental epistemic principles" (Douven and Romeijn, 2009, 16) (for the Bradley quote see Bradley, 2005, 362).

Resisting this trend, E.T. Jaynes derisively talks about the statistician-client relationship as one between a doctor and her patient. "The collection of all the logically unrelated medicines and treatments," Jaynes explains, "that a sick patient might need [is] too large for anyone but a dedicated professional to learn" (Jaynes and Bretthorst, 2003, 492.). To undercut this dependency, Jaynes suggests that

> the scientist who has learned the simple, unified principles of inference expounded here would not consult a statistician for advice because he can now work out the details of his specific data analysis problem for himself, and if necessary write a new computer program, in less time than it would take to read about them in a book or hire it done by a statistician. (Jaynes and Bretthorst, 2003, 506.)

Diaconis and Zabell warn "that any claims to the effect that maximum-entropy revision is the only correct route to probability revision should be viewed with considerable caution" (Diaconis and Zabell, 1982, 829). "Great caution" (Howson and Franklin, 1994, 456) is also what Colin Howson and Allan Franklin advise about the claim that the posterior probabilities provided by MAXENT are as like the prior probabilities as it is possible to be given the constraints imposed by the data.

The promise of PME is that it combines Jaynes' simple, unified principle with a sophisticated formal theory to confirm that the principle reliably works. Where there is inconsistency between epistemic intuitions and information theory, there is hope for deeper explanations. There is also the undeniable appeal of eclecticism and doubts that one size will fit all. The proof of the pudding will be in the eating, and so we will have a detailed look at what the inconsistencies with epistemic intuition and their solutions may be.

1.2 Counterexamples and Conceptual Problems

Although the role of PME in probability kinematics as a whole will be the scope of my work, I will pay particular attention to a problem which has stymied its acceptance by epistemologists: updating or conditioning on conditionals. Two counterexamples, Bas van Fraasen's *Judy Benjamin* and Carl Wagner's *Linguist* problem, are specifically based on updating given an observation expressed as a conditional and on PME's alleged failure to update on such observations in keeping with epistemic intuitions.

> **Example 1: Deadpool.** My 14-year-old wants to watch the movie DEADPOOL. We, the parents, disapprove. He argues that his friend James may be allowed to watch

> the movie. James' parents are friends with us. They have decided to watch the movie alone first in order to assess if it is appropriate for a 14-year-old. We ask them to share their assessment with us. Let Sherry be representative for us and Charles for James' parents. If Charles gives a positive report (movie is appropriate for 14-year-olds), then in my son's assessment the movie is appropriate for 14-year-olds. In Sherry's assessment, however, the movie remains inappropriate and Charles' reputation as assessor is compromised.

In this example, my son is epistemically entrenched that Charles is a trustworthy assessor; Sherry is epistemically entrenched that *Deadpool* is inappropriate for 14-year-olds. In terms of probabilities, what the agent learns from a specification of $P(B|A)$ depends not only on the prior probabilities but also on the epistemic entrenchment of the agent. Sometimes, the agent is so committed to B that learning a high conditional probability dramatically lowers the probability of A but leaves the posterior for B unscathed. By contrast, another agent varies $P(B)$ with what they learn about $P(B|A)$ instead of $P(A)$. More epistemically minded examples are in examples 39 and 40 on page 172. Igor Douven and Jan-Willem Romeijn propose that Hellinger's distance measure is the appropriate measure for epistemic entrenchment, so that an agent is not committed to falling on one or the other side of the extremes, but could also be epistemically leaning in one or the other direction (see Douven and Romeijn, 2009, 12ff).

Epistemic entrenchment is problematic for PME because PME has no mechanism for it. As will become clear in chapter 7, PME chooses a middling distribution between the extremes of epistemic entrenchment—this will be interpreted as problematic because in the *Judy Benjamin* case the intuition supposedly is one-sided with respect to epistemic entrenchment. Epistemic entrenchment updates on conditionals by assuming a second tier of commitment to propositions beneath the primary tier of quantitative degrees of uncertainty of belief such as probabilities (or other ways of representing uncertainty, such as ranks).

PME conceptualizes this second tier differently and is consequently at odds with the voluminous recent literature on epistemic entrenchment (for example Skyrms,

who considers higher order probabilities in Skyrms, 1986, 238; but also Hempel and Oppenheim, 1945, 114; Domotor, 1980; Earman, 1992, 52; Gillies, 2000, 21). This issue is related to a dilemma that I develop in chapter 6 for indeterminate credal states: they cannot perform the double task of representing both epistemic uncertainty and all relevant features of the evidence. The dilemma in this case, however, may be directed at PME: can it keep track of both the prior probabilities and of the epistemic entrenchments of the agent? If not, then there is a need for higher order probabilities or another way to record the entrenchments.

My line of argument emphasizes a strict dividing line between the epistemic and the evidential. Epistemic states, such as epistemic entrenchment, are not evidence and as such do not influence updating. At most, they influence how the relatively prior probabilities are generated. If the entrenchment is based on evidence, then that evidence has its usual bearing on the posterior probabilities. Higher-order probabilities are an attempt to locate more of the updating action in the realm of epistemic states. PME resists the shifting of weight from the evidential to the epistemic, just as it resists the full employment theorem.

A large part of my project is to address and defend PME's performance with respect to conditionals, both conceptually and with a view to threatening counterexamples. One issue that comes to the forefront when addressing update on conditionals is whether a rational agent has sharp credences. The rejection of indeterminate credal states used to be Bayesian dogma, but recent years have seen a substantial amount of work by Bayesians who defend and require indeterminate credal states for rational agents. I will show, using Wagner's counterexample to the PME, that it is indeed inconsistent to embrace both the PME and an affirmative attitude towards indeterminate credal states for rational agents. Wagner's counterexample only works because he implicitly assumes indeterminacy. For many contemporary Bayesians who accept indeterminacy Wagner's argument is sound, even though it is enthymematic. A defence of PME must include an independent argument against indeterminacy, which I will provide in chapter 6.

In summary, my thesis is that PME is defensible against all counterexamples and conceptual issues raised so far as a generally valid objective updating method in

probability kinematics. PME operates on the basis of a simple principle of ampliative reasoning (ampliative in the sense that the principle narrows the field of logical possibilities): when updating your probabilities, waste no useful information and do not gain information unless the evidence compels you to gain it (see Jaynes, 1988, 280; van Fraassen 1986, 376; Zellner, 1988, 278; and Klir, 2006, 356). The principle comes with its own formal apparatus (not unlike probability theory itself): Shannon's information entropy, the Kullback-Leibler divergence, the use of Lagrange multipliers, and the sometimes intricate, sometimes straightforward relationship between information and probability.

Once the counterexamples are out of the way, the more serious conceptual issues loom. On the one hand, there are powerful conceptual arguments affirming the special status of PME. Shore and Johnston, who use the axiomatic strategy of Cox's theorem in probability kinematics, show that relatively intuitive axioms only leave us with PME to the exclusion of all other objective updating methods. Van Fraassen, R.I.G. Hughes, and Gilbert Harman's MUD method, for example, or their maximum transition probability method from quantum mechanics both fulfill their five requirements (see van Fraassen et al., 1986), but do not fulfill Shore and Johnston's axioms. Neither does Uffink's more general class of inference rules, which maximize the so-called Rényi entropies, but Uffink argues that Shore and Johnston's axioms rest on unreasonably strong assumptions (see Uffink, 1995). Caticha and Giffin counter that Skilling's method of induction (see Skilling, 1988) and Jaynes' empirical results in statistical mechanics and thermodynamics imply the uniqueness of Shannon's information entropy over rival entropies.

PME seamlessly generalizes standard conditioning and Jeffrey's rule where they are applicable (see Caticha and Giffin, 2006). It underlies the entropy concentration phenomenon described in Jaynes' standard work *Probability Theory: the Logic of Science*, which also contains a sustained conceptual defence of PME and its underlying logical interpretation of probabilities. Entropy concentration refers to the unique property of the PME solution to have other distributions which obey the affine constraint cluster around it. When used to make predictions whose quality is measured by a logarithmic score function, posterior probabilities provided by PME result in

minimax optimal decisions (see Topsøe, 1979; Walley, 1991; Grünwald, 2000) so that by a logarithmic scoring rule these posterior probabilities are in some sense optimal.

Jeff Paris has investigated different belief functions (probabilities, Dempster-Shafer, and truth-functional, see Paris, 2006) from a mathematical perspective and come to the conclusion that given certain assumptions about the constraints that experience normally imposes (we will have to examine their relationship to the affine constraints assumed by PME), if a belief function is a probability function, only minimum cross entropy belief revision satisfies a host of desiderata (continuity, equivalence, irrelevant information, open-mindedness, renaming, obstinacy, relativization, and independence) while competitors fail on multiple counts (see Paris and Vencovská, 1990).

Belief revision literature, however, has in the last twenty years turned its attention to the AGM paradigm (named after Carlos Alchourrón, Peter Gärdenfors, and David Makinson), which operates on the basis of fallible beliefs and their logical relationships. There are two different epistemic dimensions, to use Henry Kyburg's expression: one where doxastic states are cashed out in terms of fallible beliefs which move in and out of belief sets; and another where 'beliefs' are reflections of a more deeply rooted, graded notion of uncertainty.

Jeffrey with his radical probabilism pursues a project of epistemological monism (see Jeffrey, 1965) which would reduce beliefs to probabilities, while Spohn and Maher seek reconciliation between the two dimensions, showing how fallible full beliefs are epistemologically necessary and how the formal structure of the two dimensions reveals many shared features so that in the end they have more in common than what separates them (see Spohn, 2012, 201 and Maher, 1993).

I am sympathetic to Spohn's and Maher's projects. PME can contribute to it in the sense of clarifying the dividing line between epistemic states and evidential input. The belief revision literature, the higher order probabilities approach, and the investigation of epistemic entrenchment all emphasize the additional influence that epistemic states have in the updating process over and above the evidence. Information theory does not lend itself to these considerations and tends to give recommendations strictly based on evidential constraints, minimizing information

gain. It may be helpful to have this clarity in partial belief epistemology before embarking on a reconciliation project with full belief epistemology.

There is a 'docta ignorantia' paradox in partial belief epistemology. While using partial beliefs rather than full beliefs at first suggests greater modesty, partial beliefs present themselves as containing more information than full beliefs. It appears that a person with a full belief in A and $\neg C$ and a non-commitment to either B or $\neg B$ is informationally more modest than a partial believer who has credences of $Cr(A) = 0.86, Cr(C) = 0.12, Cr(B) = 0.47$. Greater epistemic modesty translates into informational immodesty of doxastic states. Some Bayesians seek to remedy this paradox by introducing indeterminate credal states. I will show in chapter 5 that this strategy is inconsistent with PME. The problem with epistemic modesty expressed in imprecision is that it assumes that representation of uncertainty is just another belief. If I have a sharp credence, I have one particular belief; if I accept imprecision, I am non-committal about a range of different beliefs, which on the face of it sounds like epistemic modesty. The semantics of uncertainty, however, is such that representing uncertainty is not another belief. Information theory will help to be more clear about the representation of epistemic states and the role of epistemic states versus evidential input.

1.3 Dissertation Outline

Despite the potholes in the historical development of the PME, on account of its unifying features, its simple and intuitive foundations, and its formal success it deserves more attention in the field of belief revision and probability kinematics. PME is the single principle which can hold things together over vast stretches of epistemological terrain (intuitions, formal consistency, axiomatization, case management) and calls into question the scholarly consensus that such a principle is not needed.

Making this case, I proceed as follows. The first three chapters are introductory: preface; introduction; and a formal introduction to PME. In chapter 4, I address the geometry of reason used by defenders of the epistemic utility approach to Bayesian epistemology to justify the foundational Bayesian tenets of probabilism and standard

conditioning. The geometry of reason is the view that the underlying topology for credence functions is a metric space, on the basis of which axioms and theorems of epistemic utility for partial beliefs are formulated. It implies that Jeffrey conditioning, which is implied by PME, must cede to an alternative form of conditioning. This alternative form of conditioning fails a long list of plausible expectations and implies unacceptable results in certain cases.

One solution to this problem is to reject the geometry of reason and accept information theory in its stead. Information theory comes fully equipped with an axiomatic approach which covers probabilism, standard conditioning, and Jeffrey conditioning. It is not based on an underlying topology of a metric space, but uses non-commutative divergences instead of a symmetric distance measure. I show that information theory, despite initial promise, also fails to accommodate basic epistemic intuitions. The chapter, by bringing the alternative of information theory to the forefront, does therefore not clean up the mess but rather provides its initial cartography.

With information theory established as a candidate with both promise and areas of concern, but superior to the geometry of reason, chapters 5 and 6 address the question of indeterminate credal states. I begin with a problem that Wagner has identified for PME, the *Linguist* problem. Chapter 5 identifies the enthymematic nature of the problem: it is not mentioned that one of its premises is the acceptance of indeterminate credal states. Wagner's reasoning is valid, therefore one must either accept the conclusion and reject PME, or reject indeterminate credal states. I dedicate one whole chapter, chapter 6, to an independent defence of mandatory sharp credences for rational agents, thereby rejecting indeterminate credal states.

In chapter 5, the story goes as follows: when we come to know a conditional, we cannot straightforwardly apply Jeffrey conditioning to gain an updated probability distribution. Carl Wagner has proposed a natural generalization of Jeffrey conditioning to accommodate this case (Wagner conditioning). The generalization rests on an ad hoc but plausible intuition (W), leaving ratios of partial beliefs alone in updating when they are not affected by evidential constraints. Wagner shows how PME disagrees with intuition (W) and therefore considers PME to be undermined.

Chapter 5 presents a natural generalization of Wagner conditioning which is derived from PME and implied by it. PME is therefore not only consistent with (W), it seamlessly and elegantly generalizes it (just as it generalizes standard conditioning and Jeffrey conditioning).

Wagner's inconsistency result for (W) and PME is instructive. It rests on the assumption (I) that the credences of a rational agent may be indeterminate. While many Bayesians now hold (I) it is difficult to articulate PME on its basis because to date there is no satisfactory proposal how to measure indeterminate probability distributions in terms of information theory (see the latter part of section 6.2). Most, if not all, advocates of PME resist (I). If they did not they would be vulnerable to Wagner's inconsistency result. Wagner has therefore not, as he believes, undermined PME but only demonstrated that advocates of PME must accept that rational agents have sharp credences. Majerník shows that PME provides unique and plausible marginal probabilities, given conditional probabilities. The obverse problem posed in chapter 5 is whether PME also provides such conditional probabilities, given certain marginal probabilities. The theorem developed to solve the obverse Majerník problem demonstrates that in the special case introduced by Wagner PME does not contradict (W) (which subsequently I also call Jeffrey's Updating Principle or JUP), but elegantly generalizes it and offers a more integrated approach to probability updating.

Chapter 5 has the effect of welding together PME and a rejection of indeterminacy for credal states. Bayesians who permit indeterminate credal states will have trouble accommodating Wagner's result, while those defending PME find themselves in the position of having to defend sharp credences as well. Chapter 6 makes the case that there is integrity in a joint defence of PME and sharp credences. Many Bayesian epistemologists now accept that it is not necessary for a rational agent to hold sharp credences. There are various compelling formal theories how such a non-traditional view of credences can accommodate decision making and updating. They are motivated by a common complaint: that sharp credences can fail to represent incomplete evidence and exaggerate the information contained in it. Indeterminate credal states, the alternative to sharp credences, face challenges as well: they are vulnerable to dilation and under certain conditions do not permit learning. Chap-

ter 6 focuses on two concessions that Thomas Augustin and James Joyce make to address these challenges. The concessions undermine the original case for indeterminate credal states. I use both conceptual arguments and hands-on examples to argue that rational agents always have sharp credences.

In chapter 7, I address a counterexample that opponents of objective methods to determine updated probabilities prominently cite to undermine the generality of PME. This problem, van Fraassen's *Judy Benjamin* case, is frequently used as the knock-down argument against PME. Chapter 7 shows that an intuitive approach to Judy Benjamin's case supports PME. This is surprising because based on independence assumptions the anticipated result is that it would support the opponents. It also demonstrates that opponents improperly apply independence assumptions to the problem. Chapter 7 addresses the issue of epistemic entrenchment at length.

Chapter 2

The Principle of Maximum Entropy

2.1 Inductive Logic

David Hume poses one of the fundamental questions for the philosophy of science, the problem of induction. There is no deductive justification that induction works, as the observations which serve as a basis for inductive inference are not sufficient to make an argument for the inductive conclusion deductively valid. An inductive justification would beg the question. The late 19th and the 20th century bring us two responses to the problem of induction relevant to our project: (i) Bayesian epistemology and the subjective interpretation of probability focuses attention on uncertainty and beliefs of agents, rather than measuring frequencies or hypothesizing about objective probabilities in the world (John Maynard Keynes, Harold Jeffreys, Bruno de Finetti, Frank Ramsey; against, for example, R.A. Fisher and Karl Popper). (ii) Philosophers of science argue that some difficult-to-nail-down principle (indifference, simplicity, laziness, symmetry, entropy) justifies entertaining certain hypotheses more seriously than others, even though more than one of them may be compatible with experience (Ernst Mach, Carnap).

Two pioneers of Bayesian epistemology and subjectivism are Harold Jeffreys (see Jeffreys, 1931, and Jeffreys, 1939) and Bruno de Finetti (see de Finetti, 1931 and 1937). They personify a divide in the camp of subjectivists about probabilities. While de Finetti insists that any probability distribution could be rational for an agent to hold as long as it obeys the axioms of probability theory, Jeffreys considers probability theory to be an inductive logic with rules, resembling the rules of deduc-

tive logic, about the choice of prior and posterior probabilities. While both agree on subjectivism in the sense that probabilities reflect an agent's uncertainty (or, in Jeffreys' case, more properly a lack of information), they disagree on the subjectivist versus objectivist interpretation of how these probabilities are chosen by a rational agent (or, in Jeffreys' case, more properly by the rules of an inductive logic—as even maximally rational agents may not be able to implement them). The logical interpretation of probabilities begins with John Maynard Keynes (see Keynes, 1921), but soon turns into a fringe position with Harold Jeffreys (for example Jeffreys, 1931) and E.T. Jaynes (for example Jaynes and Bretthorst, 2003) as advocates who are standardly invoked for refutation. Carnap develops his own brand of a logical interpretation, which in some ways is analogous to his earlier conventionalism in geometry. For Carnap, there is a formal structure for partial beliefs and updating procedures, but within the formal structure there are flexibilities which allow the system to be adapted to the use to which it is put (see Carnap, 1952).

The problem in part is that the logical interpretation cannot get off the ground with plausible rules about how to choose absolutely prior probabilities. No one is able to overcome the problem of transformation invariance for the principle of indifference (consider Bertrand's paradox, see Paris, 2006, 71f), not even E.T. Jaynes (for his attempts see Jaynes, 1973; for a critical response see Howson and Urbach, 2006, 285 and Gillies, 2000, 48).

One especially intractable problem for the principle of indifference is Ludwig von Mises' water/wine paradox:

> **Example 2: The Water/Wine Paradox.** There is a certain quantity of liquid. All that we know about the liquid is that it is composed entirely of wine and water and that the ratio of wine to water is between 1/3 and 3. What is the probability that the ratio of wine to water is less than or equal to 2?

There are two ways to answer this question which are inconsistent with each other (see von Mises, 1964, 161). According to van Fraassen, example 2 shows why we should "regard it as clearly settled now that probability is not uniquely assignable on the basis of a principle of indifference" (van Fraassen, 1989, 292). Van Fraassen

goes on to claim that the paradox signals the ultimate defeat of the principle of indifference, nullifying the efforts undertaken by Poincaré and Jaynes in solving other Bertrand paradoxes (see Mikkelson, 2004, 137; and Gyenis and Rédei, 2015). Donald Gillies calls von Mises' paradox a "severe, perhaps in itself fatal, blow" to Keynes' logical theory of probability (see Gillies, 2000, 43). De Finetti's subjectivism is an elegant solution to this problem and marginalizes the logical theory.

While Jaynes throws up his hands over von Mises' paradox, despite the success he lands addressing Bertrand's paradox (see Jaynes, 1973), Jeffrey Mikkelson has recently suggested a promising solution to von Mises' paradox (see Mikkelson, 2004). There may still be hope for an objectivist approach to absolutely prior probabilities. Nevertheless, my dissertation remains agnostic about this problem. The domain of my project is probability kinematics. Relatively prior probabilities are assumed, and their priority only refers to the fact that they are prior to the posterior probabilities (one may call these distributions or densities antedata and postdata rather than prior and posterior, to avoid confusion) and not necessarily prior to earlier evidence.

This raises a conceptual problem: why would anybody be interested in a defence of objectivism in probability kinematics when the sense is that objectivism has failed about absolutely prior probabilities? My intuition is in line with Keynes, who maintains that all probabilities are conditional: "No proposition is in itself either probable or improbable, just as no place can be intrinsically distant; and the probability of the same statement varies with the evidence presented, which is, as it were, its origin of reference" (see Keynes, 1909, chapter 1).

The problem of absolutely prior probabilities is therefore moot, and it becomes clear that 'objectivism' is not really what we are advocating. Jaynes, who is initially interested in objectivism about absolutely prior probabilities as well, seems to have come around to this position when in his last work *Probability Theory: The Logic of Science* he formally introduces probabilities as conditional probabilities (and later asserts that "one man's prior probability is another man's posterior probability," see Jaynes and Bretthorst, 2003, 89). Prior probabilities in this dissertation are forever anterior, never original, as Roland Barthes puts it poetically in "The Death of the Author." Alan Hájek also considers probability the primitive notion and

unconditional probability the derivative notion, citing a list of supporters (see Hájek, 2003, 315).

The claim that all information-sharing minds ought to distribute their belief in the same way could be called Donkin's objectivism after the 19th century mathematician William Fishburn Donkin (see Zabell, 2005, 23). The logic of induction in this dissertation resists Donkin's objectivism and only provides rules for how to proceed from one probability distribution to another, given certain evidence. It is a logic without initial point (compare it perhaps to coordinate-free synthetic geometry with its affine spaces). Just as in deductive logic, we may come to a tentative and voluntary agreement on a set of rules and presuppositions and then go part of the way together. (For a more systematic analogy between deductive inference and probabilistic inference see Walley, 1991, 485; see also Kaplan, 2010, 43, for an interesting take on this; as well as Huttegger, 2015, for a more systematic reconciliation project between objectivism and resistance to Donkin's austere version of objectivism.) L.J. Savage insists in his *Foundations of Statistics* that "the criteria incorporated in the personalistic view do not guarantee agreement on all questions among all honest and freely communicating people, even in principle" (Savage, 1954, 67). My defence of PME is not inconsistent with this principle.

Bayesian probability kinematics rests on the idea that there are not only static norms about the probabilities of a rational agent, but also dynamic norms. The rational agent is not only constrained by the probability axioms, but also by standard conditioning as she adapts her probabilities to incoming evidence. Paul Teller presents a diachronic Dutch-book argument for standard conditioning (see Teller, 1973; Teller, 1976), akin to de Finetti's more widely accepted synchronic Dutch-book argument (for detractors of the synchronic Dutch-book argument see Seidenfeld et al., 1990; Foley, 1993, §4.4; and more recently Rowbottom, 2007; for a defence see Skyrms, 1987a). Brad Armendt expands Teller's argument for Jeffrey conditioning (see Armendt, 1980). In contrast to the synchronic argument, however, there is considerable opposition to the diachronic Dutch-book argument (see Hacking, 1967; Levi, 1987; and Maher, 1992). Colin Howson and Peter Urbach make the argument that standard conditioning as a diachronic norm of updating is inconsistent (see Howson and

Urbach, 2006, 81f).

An alternate route to justifying subjective degrees of belief and their probabilistic nature is to use Ramsey's approach of providing a representation theorem. Representation theorems make rationality assumptions for preferences such as transitivity (the standard reference work is still Sen, 1971) and derive from them a probability and a utility function which are unique up to acceptable transformations. Ramsey only furnishes a sketch of how this can be done. The first fully formed representation theorem is by Leonard Savage (see Savage, 1954); but soon Jeffrey notes that its assumptions are too strong. Based on mathematical work by Ethan Bolker (see Bolker, 1966; and a summary for philosophers in Bolker, 1967), Jeffrey provides a representation theorem with weaker assumptions (in Jeffrey, 1978). Since then, representation theorems have proliferated (there is, for example, a representation theorem for an acceptance-based belief function in Maher, 1993, and one for decision theory in Joyce, 1999). They are formally more complex than Dutch-book arguments, but well worth the effort because they make less controversial assumptions.

2.2 Information Theory and the Principle of Maximum Entropy

Before I can articulate my version of PME, I need to clarify what I mean by synchronic and diachronic constraints, especially in light of my comments about ignorance priors. I have been adamant not to touch absolutely prior probabilities in this dissertation and to limit myself to normativity about updating when relatively prior probabilities are already in place. This commitment is not arbitrary: I consider probabilities to be essentially conditional, not unconditional. They do not appear ex nihilo. I have no good answer at this point where these conditional probabilities come from—that is a valid question, but not the one I am addressing here. Some sort of subjectivism will have to do. It is important to me, however, and for similar reasons, that rational agents have subjective probabilities for the events they consider. The idea of indeterminate or imprecise probabilities is as suspicious to me as the idea of ignorance

priors. Both strike me as inappropriately metaphysical. In the case of indeterminate probabilities, it is as if our uncertain doxastic states successfully referred to some sort of partial truth in nature, such as frequencies; in the case of ignorance priors, it is as if nature was obligated to have each experiment correspond uniquely to a privileged set-up, for example in Bertrand's paradox, so that the epistemic agent can then apply the principle of indifference. Chapter 6 will delineate more formal reasons for the suspicion about indeterminacy.

Yet, when I articulate PME, I make reference to synchronic constraints. The diachronic constraints are not a problem: they result for example in the conditioning that is characteristic of probability update, where relatively prior probabilities are assumed. Synchronic constraints, by contrast, presuppose that a set of events currently has no subjective probabilities assigned. This appears to conflict with my Laplacean realism, which stays away from ignorance priors and Donkin's objectivism, but is committed to sharp probabilities for all events under consideration. It is the disclaimer 'under consideration' that will give me the leverage for synchronic constraints.

As throughout this dissertation, there is a finite event space Ω with $|\Omega| = n$. An agent's subjective probabilities distribution P is defined on a subset S of $\mathcal{P}(\Omega)$, the powerset of Ω. What I will allow, while retaining the out-of-bounds rule for ignorance priors and indeterminate probabilities, is domain expansion for these probability distributions. The simplest example is negation: if S is not an algebra, sometimes the probability distribution can be expanded by setting $P(-X) = 1 - P(X)$. $-X$ is the complement of X (for more details on the model see section 3.1). In most cases it makes sense to require that S be an algebra, i.e. that it is closed with respect to complements and intersections and contains $\emptyset$.

The example of a non-algebra, however, illustrates that sometimes we may want to expand the domain of a subjective probability distribution to events that come under consideration, for example the negation of a proposition (the negation of a proposition corresponds to the complement of an event in an isomorphism between propositions and events). This must not be done by introducing ignorance priors, but it can be done by having synchronic constraints for logically related propositions. It

is a synchronic constraint, for example, that the probability assigned to the negation of a proposition be such that P is still a probability distribution, therefore $P(-X) = 1 - P(X)$.

Here is another example that I will need in chapter 5. While events X and Y may be under consideration, their intersection may not be. As with negation, this will only happen if the original S is not an algebra. In contrast to negation, however, many different joint probability distributions are compatible with the given marginal probability distribution. The synchronic constraint to fill in the joint probabilities using the probability axioms does not yield a unique distribution. This is where PME makes a controversial recommendation: fill in the joint probabilities as if X and Y were independent, unless you have information to the contrary. I will show how this recommendation derives from PME, explain its relation to independence, and defend it at various points in my dissertation.

My version of PME goes as follows:

> If a rational agent is subject to constraints describing, but not uniquely identifying her doxastic state in terms of coherent probability distributions, then the doxastic state of the agent is characterized by the unique probability distribution identified by information theory, if it exists. If the constraints are synchronic, the principle of maximum entropy is used. If they are diachronic, the principle of minimum cross-entropy is used.

Here are four examples to illuminate PME. For simplicity, I am assuming that only one of the following three candidates will be the eventual Republican nominee for the presidential election in the US in 2016: Donald Trump, Ted Cruz, and Marco Rubio. In the following examples, a computer program called Rasputin which has no information but the information provided in the example is learning how to provide rational betting odds for forced bets. I am using betting odds from a computer instead of the doxastic state of a rational agent merely for illustration. Setting a betting threshold at p means that, if forced to either accept a bet on proposition X at a cost of $\hat{p}$ or on $\neg X$ at a cost of $1 - \hat{p}$, Rasputin will choose the former bet if

$p > \hat{p}$ and the latter bet if $p \leq \hat{p}$. Bets are always placed such that the prize is \$1 if the proposition turns out to be true and nothing if it turns out to be false.

Example 3: Trump Standard Conditioning. A prediction market like `www.pivit.io` takes the following bets: 66 cents on the dollar for Donald Trump to be the Republican nominee. 21 cents on the dollar for Ted Cruz. 13 cents on the dollar for Marco Rubio. Cruz announces (reliably) that he will drop out of the race if he loses Texas. Rasputin sets the betting threshold for Trump at 83.544 and for Rubio at 16.456 cents on the dollar in case Cruz loses Texas (this is a conditional bet which is called off with the cost of the bet reimbursed if the antecedent does not hold true).

Example 4: Trump Jeffrey Conditioning. Given the odds in example 3, the prediction market also takes 97 cents on the dollar bets that if Rubio loses Florida, he will not be the Republican nominee (another conditional bet). Rasputin sets the betting threshold for Trump at 73.586 and for Cruz at 23.414 cents on the dollar in case Rubio loses Florida.

Example 5: Trump Affine Constraint. Due to a transmission problem, Rasputin receives an incomplete update of current betting ratios after the primary in Michigan. Given the odds in example 3, Rasputin can only add the information that if either Trump or Cruz win the nomination (another conditional bet), the bets on Trump are 85 cents on the dollar. Rasputin now has the option of keeping Rubio's odds constant at 13 cents or changing them in accordance with the principle of minimum cross-entropy (this is the *Judy Benjamin* problem). Rasputin applies the principle of minimum cross-entropy and sets the betting threshold for Rubio at 13.289 cents on the dollar. I will explain in subsection 3.2.3 how to calculate this number, using the constraint rule for cross-entropy.

Example 6: Trump Marginal Probabilities. Rasputin has the betting odds for Trump, Cruz, and Rubio on the Republican side (66:21:13) and the betting odds for Clinton and Sanders on the Democratic side (82:18). Rasputin needs to set betting thresholds for forced bets on the joint events, such as 'Cruz will win the Republican

> nomination and Clinton will win the Democratic nomination'. Given the marginal probabilities, many combinations of joint probabilities are probabilistically coherent. Rasputin chooses a unique set of joint probabilities by multiplying the marginal probabilities as in the following table because they have the highest entropy of those that are probabilistically coherent.

$\wedge$	Trump	Cruz	Rubio	
Clinton	0.5412	0.1722	0.1066	0.82
Sanders	0.1188	0.0378	0.0234	0.18
	0.66	0.21	0.13	1.00

As for historical development, when Jaynes introduces his version of PME (see Jaynes, 1957a; 1957b), he is less indebted to the philosophy of science project of giving an account of semantic information (as in Carnap and Bar-Hillel, 1952; 1953) than to Claude Shannon's mathematical theory of information and communication. Shannon identifies information entropy with a numerical measure of a probability distribution fulfilling certain requirements (for example, that the measure is additive over independent sources of uncertainty). The focus is not on what information is but how we can formalize an axiomatized measure. Entropy stands for the uncertainty that is still contained in information (certainty is characterized by zero entropy).

Shannon introduces information entropy in 1948 (see Shannon, 2001), based on work done by Norbert Wiener connecting probability theory to information theory (see Wiener, 1939). Jaynes also traces his work back to Ludwig Boltzmann and Josiah Gibbs, who build the mathematical foundation of information entropy by investigating entropy in statistical mechanics (see Boltzmann, 1877; Gibbs, 1902).

For the further development of PME in probability kinematics it is important to refer to the work of Richard Jeffrey, who establishes the discipline (see Jeffrey, 1965), and Solomon Kullback, who provides the mathematical foundations of minimum cross-entropy (see Kullback, 1959). In probability kinematics, contrasted with standard conditioning, evidence is uncertain (for example, the ball drawn from an urn may have been observed only briefly and under poor lighting conditions).

Jeffrey addresses many of the conceptual problems attending probability kinematics by providing a much improved representation theorem, thereby creating a tight connection between preference theory and its relatively plausible axiomatic foundation and a probabilistic view of 'beliefs.' Jeffrey and Isaac Levi's (see Levi, 1967) debates on partial belief and acceptance (revisable and not necessarily certain full belief), which Jeffrey considered to be as opposed to each other as Dracula and Wolfman (see Jeffrey, 1970), set the stage for two 'epistemological dimensions' (Henry Kyburg's term, see Kyburg, 1995, 343), which will occupy us in detail and towards which I will take a more conciliatory approach, as far as their opposition or mutual exclusion is concerned.

Kullback's divergence relationship between probability distributions makes possible a smooth transition from synchronic arguments about absolutely prior probabilities to diachronic argument about probability kinematics (this transition is much more troublesome from the synchronic Dutch-book argument to the diachronic Dutch-book argument; for the information-theoretic virtues of the Kullback-Leibler divergence see Kullback and Leibler, 1951; Seidenfeld, 1986, 262ff; Guiaşu, 1977, 308ff).

Jaynes' project of probability as a logic of science is orginally conceived to provide objective absolutely prior probabilities by using PME, rather than to provide objective posterior probabilities, given relatively prior probabilities. It is, however, easy to turn PME into a method of probability kinematics using the Kullback-Leibler divergence. Jaynes presents this method in 1978 at an MIT conference under the title "Where Do We Stand on Maximum Entropy?" (see Jaynes, 1978), where he explains the *Brandeis* problem and demonstrates the use of Lagrange multipliers in probability kinematics.

Ariel Caticha and Adom Giffin have recently demonstrated, using Lagrange multipliers, that PME seamlessly generalizes standard conditioning (see Caticha and Giffin, 2006). Many others, however, think that in one way or another PME is inconsistent with standard conditioning, to the detriment of PME (see Seidenfeld, 1979, 432f; Shimony, 1985; van Fraassen, 1993, 288ff; Uffink, 1995, 14; and Howson and Urbach, 2006, 278); Jon Williamson believes so, too, but to the detriment of standard

conditioning (see Williamson, 2011).

Arnold Zellner proves that standard conditioning as a diachronic updating rule (Bayes' theorem) is the "optimal information processing rule" (Zellner, 1988, 278), also using Lagrange multipliers (Jaynes already has this hunch in Jaynes, 1988, 280; see also van Fraassen et al., 1986, 376). Standard conditioning is neither inefficient (using a suitable information metric), diminishing the output information compared to the input information, nor does it add extraneous information. This is just the simple conceptual idea behind PME, although PME only requires optimality, not full efficiency. Full efficiency implies optimality, therefore standard conditioning fulfills PME.

Once PME is formally well-defined and its scope established (for the latter, Imre Csiszár's work on affine constraints is important, see Csiszár, 1967), its virtues come to the foreground. While Richard Cox (see Cox, 1946) and E.T. Jaynes defend the idea of probability as a formal system of logic, John Shore and Rodney Johnson provide the necessary detail to establish the uniqueness of PME in meeting intuitively compelling axioms (see Shore and Johnson, 1980).

2.3 Criticism

By then, however, an avalanche of criticism against PME as an objective updating method had been launched. Papers by Abner Shimony (see Friedman and Shimony, 1971; Dias and Shimony, 1981; Shimony, 1993) note the following problem.

> **Example 7: Shimony's Lagrange Multiplier.** You mail-order a four-sided die whose spots are 1, 2, 3, and 6 respectively (I am using this example leaning on Jaynes' *Brandeis Dice* problem). The die is mailed to a lab, where a graduate student takes it out of its package and rolls it thousands of times. After writing down the long-run average μ, she rolls it one more time and reports the result of the last die roll to you. X_k is the proposition that the die roll comes up k spots on the last die roll. Before you hear the result, you have no idea what the die's bias is. Your relative priors happen to be equiprobabilities, so $P(X_k) = 1/4$ for $k = 1, 2, 3, 6$.

Let Y_z be the event indicating that the long-run average as recorded by the graduate student is z with $z \in [1, 6] \subset \mathbb{R}$. h is the probability density for Y_z,

$$P(a \leq z \leq b) = \int_a^b h(z)\, dz, \tag{2.1}$$

and $g_k(z)$ is the probability density for $P(X_k|Y_z)$,

$$P\left(X_k| \bigcup_{z\in[a,b]} Y_z\right) = \int_a^b g_k(z)\, dz. \tag{2.2}$$

The law of total probability gives us the following (integrals without subscript are integrals over $\mathbb{R}$):

$$\frac{1}{4} = P(X_k) = \int g_k(z)h(z)\, dz, \tag{2.3}$$

Let ν be the measure associated with the probability density h,

$$\nu(A) = P(\bigcup_{z\in A} Y_z), \tag{2.4}$$

where A is a measurable subset of $\mathbb{R}$. In our example, $\nu(A)$ is trivially zero for $A \cap (1, 6) = \emptyset$. Then we may write for any $k = 1, 2, 3, 6$,

$$P(X_k) = \int \frac{e^{-\beta k}}{\sum_{i=1,2,3,6} e^{-\beta i}}\, d\mu \tag{2.5}$$

according to the constraint rule of cross-entropy, which will be introduced in subsection 3.2.3. There is an isomorphism φ between the long-run average z and the

Lagrange multiplier β which provides the basis for the definition of the measure

$$\mu(B) = \nu(\varphi^{-1}(B)) \text{ for any measurable } B \subseteq \mathbb{R}. \tag{2.6}$$

Simplifying (2.5) to

$$P(X_k) = \int \left(\sum_{i=1,2,3,6} e^{\beta(k-i)} \right)^{-1} d\mu = \int I(\beta)\, d\mu, \tag{2.7}$$

the derivative of the integrand $I(\beta)$ is

$$\frac{dI(\beta)}{d\beta} = -\left(\sum_{i=1,2,3,6} e^{\beta(k-i)} \right)^{-2} \sum_{i=1,2,3,6} (k-i) e^{\beta(k-i)}, \tag{2.8}$$

which vanishes at $\beta = 0$ for $k = 3$. Let us assume $k = 3$ from now on. Shimony is making sure that k is one of the possible outcomes and at the same time their mean. In example 7, I fulfill this requirement by choosing $1, 2, 3, 6$, where 3 is the mean. A regular die, whose fair mean is 3.5, would not fulfill it because 3.5 is not one of the outcomes.

Straightforward calculations show that $I(\beta)$ has a maximum at $\beta = 0$, which is the only extremum (a maximum) of $I(\beta)$. To summarize the properties of I, (i) it is extremal at zero; (ii) $I(0) = 1/4$; and (iii) $\int I(\beta)\, d\mu = 1/4$. These three properties imply that $\mu(\{0\}) = 1$. To see this, consider the following sequence of subsets of $\mathbb{R}$:

$$L_n = \left\{ \beta : I(\beta) < \frac{1}{4} - \frac{1}{n} \right\}. \tag{2.9}$$

Then,

$$\frac{1}{4} = \int I(\beta)\, dx = \int_{L_n} I(\beta)\, d\mu + \int_{\mathbb{R}\setminus L_n} I(\beta)\, d\mu \leq$$
$$\left(\frac{1}{4} - \frac{1}{n}\right) \mu(L_n) + \frac{1}{4}(1 - \mu(L_n)) = \frac{1}{4} - \frac{\mu(L_n)}{n}, \tag{2.10}$$

so $\mu(L_n) = 0$. Now by countable additivity,

$$\mu(\mathbb{R}\setminus\{0\}) = \mu\left(\bigcup_{n\in\mathbb{N}} L_n\right) = 0. \tag{2.11}$$

Before you know anything else, you ought to know with certainty that the long-run average recorded by the graduate student is 3, which is the value of z corresponding to $\beta = 0$. This conflicts with the assumption that you know nothing about the die when it arrives in the mail.

Shimony's more general description of example 7 convinces among many others Brian Skyrms that PME and its objectivism are not tenable (see Skyrms, 1985; 1986; and 1987b). In section 2.5, I am providing literature to defend PME against the claims of Shimony's Lagrange multiplier problem. In a nutshell, I can say that it is the isomorphism φ in conjunction with (2.2) that draws Jaynes' withering criticism (see Jaynes, 1985, 134ff). While z is a legitimate statistical parameter of the distribution for X_k, β is an artefact of the mathematical process to find a posterior probability, not a conditional probability. (2.3), however, is only true for conditional probabilities. The mathematical artefact β is defined by the procedure and therefore not uncertain as the parameter z would be, even if there is a one-one correspondence between their values. z is within the domain of the prior probability distribution and its conditional probabilities, β is not.

Following Shimony's Lagrange multiplier problem, Bas van Fraassen's *Judy Benjamin* problem (see van Fraassen, 1981) deals another blow to PME in the literature, motivating Joseph Halpern (who already has reservations against Cox's theorem, see

Halpern, 1999) to reject it in his textbook on uncertainty (see Halpern, 2003). I will cover the *Judy Benjamin* problem at length in chapter 7.

Teddy Seidenfeld runs a campaign against objective updating methods in articles such as "Why I Am Not an Objective Bayesian" (see Seidenfeld, 1979; 1986). Jos Uffink takes issue with Shore and Johnson, casting doubt on the uniqueness claims of PME (see Uffink, 1995; 1996). Carl Wagner introduces a counterexample to PME (see Wagner, 1992), again, as in the *Judy Benjamin* counterexample (but in much greater generality), involving conditioning on conditionals.

In 2003, Halpern criticizes PME with the help of Peter Grünwald and the concept of 'coarsening at random,' which according to the authors demonstrates that the PME "essentially never gives the right results" (see Grünwald and Halpern, 2003, 243). A CAR condition specifies the requirements necessary to keep conditioning on a naive space consistent with conditioning on a sophisticated space. De Finetti's exchangeability is an example for a CAR condition: as long as exchangeability holds, conditioning on a simpler event space yields results that are consistent with conditioning on a more complicated event space. Grünwald and Halpern seek to show that an application of PME on a naive space is almost always inconsistent with the proper application of standard conditioning on the corresponding sophisticated space. Especially in the *Judy Benjamin* case, the fact that there is no non-trivial CAR condition for PME undermines the result that PME provides. I address this question in more detail in section 7.4.

In 2009, Douven and Romeijn write an article on the *Judy Benjamin* problem (see Douven and Romeijn, 2009) in which they ask probing questions about the compatibility of objective updating methods with epistemic entrenchment. I will address these questions in section 7.3.

Malcolm Forster and Elliott Sober's attack on Bayesian epistemology using Akaike's Information Criterion is articulated in the 1990s (see Forster and Sober, 1994) but reverberates well into the next decade (Howson and Urbach call the authors the 'scourges of Bayesianism'). Because the attack concerns Bayesian methodology as a whole, it is not within our purview to defend PME against it (for a defence of Bayesianism see Howson and Urbach, 2006, 292ff), but it deserves mention for its

direct reference to information as a criterion for inference and provides an interesting point of comparison for maximum entropy.

Another criticism which affects both the weaker Bayesian claim for standard conditioning and the stronger PME is its purported excessive apriorism, i.e. the concern that the agent can never really move away from beliefs once formed—and that those beliefs always need to be fully formed all the time. It can be found as early as 1945 in Carl Hempel (see Hempel and Oppenheim, 1945, 107) and is raised again as late as 2005 by James Joyce (see Joyce, 2005, 170f). Peter Walley uses a similar point to criticize the Bayesian position (see Walley, 1991, 334). In *Bayes or Bust?,* excessive apriorism (among other things) leads John Earman to his tongue-in-cheek position of being a Bayesian only on Mondays, Wednesdays, and Fridays (see Earman, 1992, 1; for the detailed criticism Earman, 1992, 139f).

Gillies also has these reservations (see Gillies, 2000, 81; 84) and cites de Finetti in support,

> If you want to apply mathematics, you must act as though the measured magnitudes have precise values. This fiction is very fruitful, as everybody knows; the fact that it is only a fiction does not diminish its value as long as we bear in mind that the precision of the results will be what it will be ... to go, with the valid help of mathematics, from approximate premises to approximate conclusions, I must go by way of an exact algorithm, even though I consider it an artifice. (De Finetti, 1931, 303.)

Seidenfeld militates against objective Bayesianism in 1979 using excessive apriorism (I owe the term to him, see Seidenfeld, 1979, 414). Again, because the charge is directed at Bayesians more generally, I do not need to address it, but mention it because PME may have resources at its disposal that the more general Bayesian position lacks (for this position, see Williamson, 2011).

A more recent criticism centres on the concept of epistemic entrenchment, which I will introduce in the next section. While epistemic entrenchment has attracted attention, not least on account of some interesting formal features and the ease with which it treats updating on conditionals, information theory has difficulty accommodating it.

2.4 Acceptance versus Probabilistic Belief

Epistemic entrenchment figures prominently in the AGM literature on belief revision (for one of its founding documents see Alchourrón et al., 1985) and is based on two levels of uncertainty about a proposition: its static inclusion in belief sets on the one hand, and its dynamic behaviour under belief revision on the other hand. It is one thing, for example, to think that the probability of a coin landing heads is 1/2 and consider it fair because you have observed one hundred tosses of it, or to think that the probability of a coin landing heads is 1/2 because you know nothing about it. In the former scenario, your belief that $P(X = H) = 0.5$ is more entrenched.

Wolfgang Spohn provides an excellent overview of the interplay between Bayesian probability theory, AGM belief revision, and ranking functions (see Spohn, 2012). The extent to which PME is compatible with epistemic entrenchment and a distinction between the static and the dynamic level will be a major topic of my investigation. At first glance, PME and epistemic entrenchment are at odds, because PME operates without recourse to a second epistemic layer behind probabilities expressing uncertainty. Our conclusion is that the content of this layer is expressible in terms of evidence and is not epistemic.

For a long time, there has been unease between defenders of partial belief (such as Richard Jeffrey) and defenders of full (and usually defeasible or fallible) belief (such as Isaac Levi). This issue is viewed more pragmatically beginning in the 1990s with Patrick Maher's *Betting on Theories* (see Maher, 1993) and Wolfgang Spohn's work in several articles (later summarized in Spohn, 2012). Both authors seek to downplay the contradictory nature of these two approaches and emphasize how both are necessary and able to inform each other.

Maher argues that representation theorems are superior to Dutch-book arguments in justifying Bayesian methodology, but then distinguishes between practical utility and cognitive utility. Whereas probabilism is appropriate in the arena of acting, based on practical utility, acceptance is appropriate in the arena of asserting, based on cognitive utility. Maher then provides his own representation theorem with respect to cognitive utility, underlining the resemblance in structure between the probabilistic

and the acceptance-based approach.

In a similar vein, Spohn demonstrates the structural similarities between the two approaches using ranking theory for the acceptance-based approach. Together with the formal methods of the AGM paradigm, ranking theory delivers results that are analogous to the already well-formulated results of Bayesian epistemology. Maher and Spohn put us on the right track of reconciliation between the two epistemological dimensions, and I hope to contribute to it by showing that PME can be coherently coordinated with this reconciliation. This will only be possible if we clarify the relation that PME has to epistemic entrenchment and how it conditions on conditionals. For more detail on this see especially section 7.3.

2.5 Proposal

My interpretation of PME is an intermediate position between what I would call Jaynes' Laplacean idealism, where evidence logically prescribes unique and determinate probability distributions to be held by rational agents; and a softened version of Bayesianism exemplified by, for example, Richard Jeffrey and James Joyce (for the latter see Joyce, 2005).

I side with Jaynes in so far as I am committed to determinate prior probabilities, whether they are absolute or relative (why this is important will become clear in chapter 5). Once a rational agent considers a well-defined event space, the agent is able to assign fixed numerical probabilities to it (this ability is logical, not practical—in practice, the assignment may not be computationally feasible). This assignment is either irreducibly relative to the updating process (where conditional probabilities are conceptually foundational to unconditional probabilities, not the other way around), or it depends on logical relationships to existing probabilities which provide the necessary synchronic constraints. Because there is no objectivity in the genealogy of conditional probabilities and in the interpretation of evidence, both of which introduce elements of subjectivity, I part ways with Jaynes on objectivity.

'Humanly faced' Bayesians such as Jeffrey or Joyce claim that rational agents typically lack determinate subjective probabilities and that their opinions are char-

acterized by imprecise credal states in response to unspecific and equivocal evidence. There is a difference, however, between (1) appreciating the imprecision in interpreting observations and, in the context of probability updating, casting them into appropriate mathematical constraints for updated probability distributions, and (2) bringing to bear formal methods to probability updating which require numerically precise priors. The same is true for using calculus with imprecise measurements. The inevitable imprecision in our measurements does not attenuate the logic of using real analysis to come to conclusions about the volume of a barrel or the area of an elliptical flower bed. My project will seek to articulate these distinctions more explicitly.

My project therefore promotes what I would call Laplacean realism, contrasting it with Jaynes' Laplacean idealism (Sandy Zabell uses the less complimentary term "right-wing totalitarianism" for Jaynes' position, see Zabell, 2005, 28), but also distinguishing it from contemporary softened versions of Bayesianism such as Joyce's or Jeffrey's (Zabell's corresponding term is "left-wing dadaists," although he does not apply it to Bayesians). What is distinctive about my approach to Bayesianism is the high value I assign to the role that information theory plays within it. My contention is that information theory provides a logic for belief revision. Almost all epistemologists, who are Bayesians, currently have severe doubts that information theory can deliver on this promise, not to mention their doubts about the logical nature of belief revision (see for example Zabell, 2005, 25, where he repeatedly charges that advocates of logical probability have never successfully addressed Ramsey's "simple criticism" about how to apply observations to the logical relations of probabilities).

One way in which these doubts can be addressed is by referring them to the more general debate about the relationship between mathematics and the world. The relationship between probabilities and the events to which they are assigned is not unlike the relationship between the real numbers we assign to the things we measure and calculate and their properties in the physical world. As unsatisfying as our understanding of the relationship between formal apparatus and physical reality may be, the power, elegance, and internal consistency of the formal methods is rarely in dispute. Information theory is one such apparatus, probability theory is another.

In contemporary epistemology, their relationship is held to be at best informative of each other. Whenever there are conceptual problems or counterintuitive examples, the two come apart. I consider the relationship to be more substantial than currently assumed.

There have been promising and mathematically sophisticated attempts to define probability theory in terms of information theory (see for example Ingarden and Urbanik, 1962; Kolmogorov, 1968; Kampé de Fériet and Forte, 1967—for a detractor who calls information theory a "chapter of the general theory of probability" see Khinchin, 1957). While interesting, making one of information theory or probability theory derivative of the other is not my project. What is at the core of my project is the idea that information theory delivers the unique and across the board successful candidate for an objective updating mechanism in probability kinematics. This idea is unpopular in the literature, but the arguments on which the literature relies are not robust, neither in quantity nor in quality. There are only a handful of counterexamples, none of which are ultimately resistant to explanation by a carefully articulated version of my claim.

Carnap advises pragmatic flexibility with respect to inductive methods, although he presents only a one-dimensional parameter system of inductive methods which curbs the flexibility. On the one hand, Dias and Shimony report that Carnap's λ-continuum of inductive methods is consistent with PME only if $\lambda = \infty$ (see Dias and Shimony, 1981). This choice of inductive method is unacceptable even to Carnap, albeit allowed by the parameter system, because it gives no weight to experience (see Carnap, 1952, 37ff). On the other hand, Jaynes makes the case that PME entails Laplace's Rule of Succession ($\lambda = 2$) and thus occupies a comfortable middle position between giving all weight to experience ($\lambda = 0$, for the problems of this position see Carnap, 1952, 40ff) or none at all ($\lambda = \infty$). While Carnap's parameter system of inductive methods rests on problematic assumptions, I surmise that Dias and Shimony's assignment of λ, given PME, is erroneous, and that Jaynes' assignment is better justified (for literature see the next paragraph). Since this is an old debate, I will not resurrect it here.

In order to defend my position, I need to address three important counterexamples

which at first glance discredit PME: Shimony's Lagrange multiplier problem, van Fraassen's *Judy Benjamin* case, and Wagner's *Linguist*. For the latter two, I am confident that we can make a persuasive case for PME, based on formal features of these problems which favour PME on closer examination. For the former (Shimony), Jaynes has written a spirited rebuttal (see Jaynes, 1985, 134ff). According to Jaynes, errors in Shimony's argument have been pointed out five times (see Hobson, 1972; Tribus and Motroni, 1972; Gage and Hestenes, 1973; Jaynes, 1978; Cyranski, 1979). This does not, however, keep Brian Skyrms, Jos Uffink, and Teddy Seidenfeld from referring again to Shimony's argument in rejecting PME in the 1980s.

As mentioned before, Jeffrey with his radical probabilism pursues a project of epistemological monism (see Jeffrey, 1965) which would reduce beliefs to probabilities, while Spohn and Maher seek reconciliation between the two dimensions, showing how fallible full beliefs are epistemologically necessary and how the formal structure of the two dimensions reveals many shared features so that in the end they have more in common than what separates them (see Spohn, 2012, 201 and Maher, 1993).

In the end, our project is not about the semantics of doxastic states. We do not argue the eliminativism of beliefs in favour of probabilities; on the contrary, the belief revision literature has opened an important door for inquiry in the Bayesian dimension with its concept of epistemic entrenchment. This is a good example for the kind of cross-fertilization between the two different dimensions that Spohn has in mind, mostly in terms of formal analogies and with little worry about semantics, important as they may be. Maher provides similar parallels between the two dimensions, also with an emphasis on formal relationships, in terms of representation theorems. Pioneering papers in probabilistic versions of epistemic entrenchment are recent (see Bradley, 2005; Douven and Romeijn, 2009).

The guiding idea behind epistemic entrenchment is that once an agent is apprised of a conditional (indicative or material), she has a choice of either adjusting her credence in the antecedent or the consequent (or both). Often, the credence in the antecedent remains constant and only the credence in the consequent is adjusted (Bradley calls this 'Adams conditioning'). Douven and Romeijn give an example where the opposite is plausible and the credence in the consequent is left constant

(see Douven and Romeijn, 2009, 12). Douven and Romeijn speculate that an agent can theoretically take any position in between, and they use Hellinger's distance to represent these intermediary positions formally (see Douven and Romeijn, 2009, 14).

Bradley, Douven, and Romeijn use a notion frequently used and introduced by the AGM literature to capture formally analogous structures in probability theory. The question is how compatible the use of epistemic entrenchment in probabilistic belief revision (probability kinematics) is with PME. PME appears to assign probability distributions to events without any heed to epistemic entrenchment. The *Judy Benjamin* problem is a case in point. PME's posterior probabilities are somewhere in between the possible epistemic entrenchments, as though mediating between them, but they affix themselves to a determinate position (which in some quarters raises worries analogous to excessive apriorism).

Epistemologists are generally agreed that PME is not needed because they expect pragmatic latitude in addressing questions of belief revision. The full employment theorem prefers a wide array of updating methods to a reduction and narrowing of choices. As there is already widespread consensus that objectivism cannot provide a unique set of absolutely prior probabilities, nobody sees any reason that objectivism should succeed in probability kinematics. The problem with this position is that the wide array of updating methods systematically leads to solutions which contradict the principle that a rational agent uses relevant information and does not gain unwarranted information, and that for most constraints this principle leads to a unique solution optimally fulfilling it.

In the quest to undermine uniqueness, independence assumptions often sneak in through the back door where they are indefensible (see section 7.3). Well-posed problems are dismissed as ambiguous (e.g. the *Judy Benjamin* problem), while problems that are overdetermined may be treated as open to a whole host of solutions because they are deemed underdetermined (e.g. von Mises' water and wine paradox). Ad hoc updating methods proliferate which can often be subsumed under PME with little mathematical effort (e.g. Wagner's *Linguist* problem and its treatment in chapter 5).

Chapter 3

Changing Partial Beliefs Using Information Theory

3.1 The Model

This chapter describes in more detail how probabilities are updated using information theory. First we distinguish between full beliefs and partial beliefs. A full belief formal epistemology, such as in Spohn, 2012, provides an account of what it means for an agent to accept a proposition X, how this acceptance is quantitatively related to other beliefs, as for example in ranking theory, and how it changes with new evidence. Partial belief formal epistemology provides this account using a function that assigns to some propositions a non-negative real number representing partial belief.

In this dissertation, I am limiting myself to finite outcome spaces and finite propositional languages. Let Ω be a finite outcome space with $|\Omega| = n$. Since Ω is finite, there is no issue with ill-behaved power sets and no need to use a σ-algebra instead, so let $\mathcal{S}$ be the power set of Ω. A partial belief function B has as its domain a subset $S \subseteq \mathcal{S}$ and as its range the non-negative real numbers. Assuming probabilism, B has the following properties: (P1) $B(\Omega) = 1$; and (P2) if $X \cap Y = \emptyset$ then $B(X \cup Y) = B(X) + B(Y)$. Dempster-Shafer belief functions are an alternative to probabilistic belief functions (see Dempster, 1967). Since I will almost exclusively talk about probabilistic belief functions from now on, I will use the mathematical notation P instead of B. Instead of sets, I will sometimes apply P to propositions, assuming an isomorphism between the outcome space and a corresponding propositional language for which the proposition $\tilde{X} \vee \tilde{Y}$ corresponds

to the set $X \cup Y$, the proposition $\tilde{X} \wedge \tilde{Y}$ corresponds to the set $X \cap Y$, and the proposition $\neg\tilde{X}$ corresponds to $-X$.

There are special probabilistic partial belief functions P_Y called conditional probability functions which have the additional property that $P_Y(Y) = 1$. We usually write $P_Y(X) = P(X|Y)$. For Bayesians,

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)}. \tag{3.1}$$

Bayes' formula is a direct consequence of (3.1), provided $P(Y) \neq 0$,

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}. \tag{3.2}$$

I am not concerned here with absolutely prior beliefs, so for the rest of this dissertation the assumption is that there is a belief function P, wherever it may have come from, and that our interest is focused on how to update it to a belief function P', given some intervening evidential input. For this evidential input, I restrict myself to the consideration of constraints in the form

$$\sum_{i=1}^{n} c_{ij} P'(X_i) = b_j \tag{3.3}$$

for an $n \times k$ coefficient matrix $C = (c_{ij})$ and a k-dimensional vector b. I will give a few examples first and then explain why constraints in this form are called affine constraints and why non-affine constraints would introduce problems which are wise to bracket for the moment.

The best-known example for an affine constraint is standard conditioning. In subsection 4.2.3, example 11 tells the story of a criminal case in which Sherlock Holmes' probability distribution over three people is

$$P(E_1) = 1/3, P(E_2) = 1/2, P(E_3) = 1/6 \tag{3.4}$$

where E_1 is the proposition that the first person committed the crime. Let us say Holmes finds out that the third person did not commit the crime. The updating of $P(E_1)$ and $P(E_2)$ in light of $P(E_3) = 0$ is called standard conditioning. Holmes' evidence in this case is $k = 1, b_1 = 0$ and

$$C = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{3.5}$$

Another well-known example for an affine constraint different from standard conditioning is a Jeffrey-type updating scenario. If Holmes' first person is male and the other two persons are female, then Holmes may find out that the probability of the culprit being male is 1/2 after holding the relatively prior probabilities in (3.4). He needs to change $P(E_1) = 1/3$ to $P'(E_1) = 1/2$. In the language of (3.3), his evidence is $k = 1, b_1 = 0.5$ and

$$C = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \tag{3.6}$$

A third example of an affine constraint, different from both standard and Jeffrey-type updating scenarios, is illustrated by Jaynes' *Brandeis Dice* problem introduced in his 1962 Brandeis lectures.

> **Example 8: Brandeis Dice.** When a die is rolled, the number of spots up can have any value i in $1 \leq i \leq 6$. Suppose a die has been rolled many times and we are told only that the average number of spots up was not 3.5 as we might expect from

> an 'honest' die but 4.5. Given this information, and nothing else, what probability should we assign to i spots on the next roll? (See Jaynes, 1989, 243.)

Jaynes considers the constraint in this case to be the mean value constraint

$$\sum_{i=1}^{6} iP'(X_i) = 4.5 \tag{3.7}$$

where X_i is the proposition that i spots will be up on the next roll. In the language of (3.3), the evidence is $k = 1, b_1 = 4.5$ and

$$C = (1, 2, 3, 4, 5, 6)^{\intercal}. \tag{3.8}$$

The disjunctive normal form theorem in logic guarantees that any proposition in our finite propositional language is expressible by a conjunction of atomic propositions (corresponding to an atomic outcome $\omega \in \Omega$) and their negations. Consequently, any probabilistic belief function P is determined by its values $P(\omega)$ on the atoms of the outcome space (see Paris, 2006, 13). This means that every probabilistic belief function is uniquely represented by a point on the $n-1$-dimensional simplex $\mathbb{S}^{n-1} \subset \mathbb{R}^n$ in the following way: if $P(\omega_i) = x_i$ for $i = 1, \ldots, n$ then this probabilistic belief function corresponds to $X = (x_1, \ldots, x_n)$ in $\mathbb{R}^n$ with

$$\sum_{i=1}^{n} x_i = 1. \tag{3.9}$$

(3.9) is the identifying characteristic for points on the simplex $\mathbb{S}^{n-1}$. Statisticians usually want to model these cases using parameters. For finite outcome spaces, the model has $n-1$ parameters, for example $x_2, \ldots, x_n$. x_1 is then determined to be

$$x_1 = 1 - \sum_{i=2}^{n} x_i. \tag{3.10}$$

The statistical model, an affine constraint, and an interesting ambivalence in updating methods is illustrated in figures 3.1 and 3.2.

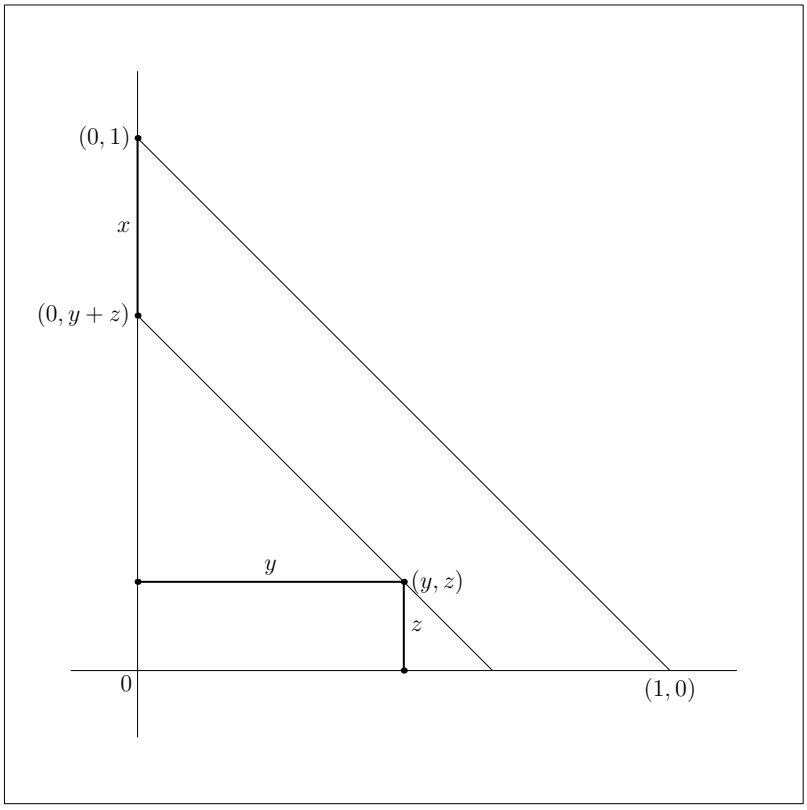

Figure 3.1: The statistical model of the 2-dimensional simplex $\mathbb{S}^2$. y and z are the parameters, whereas x is fully determined by y and z. x, y, z represent the probabilities $P(X_1), P(X_2), P(X_3)$ on an outcome space $\Omega = \{\omega_1, \omega_2, \omega_3\}$, where X_i is the proposition corresponding to the atomic outcome ω_i. The point (y, z) corresponds to the numbers in example 11.

As long as constraints of the form (3.3) are consistent, they identify a subset of the simplex $\mathbb{S}^{n-1}$ which is an affine subset. That is why we call them affine constraints.

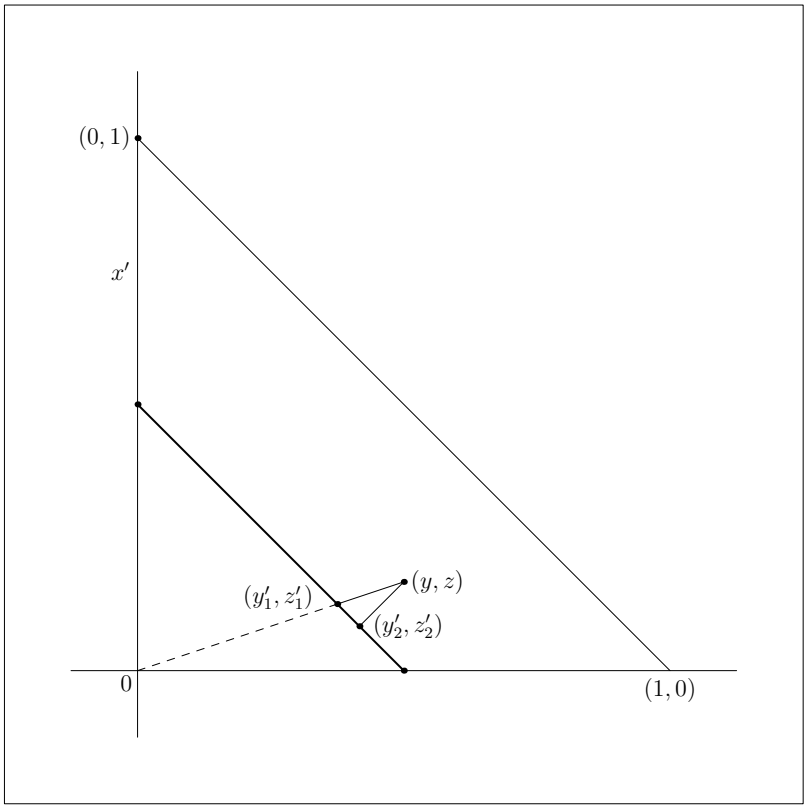

Figure 3.2: Illustration of an affine constraint. This is a Jeffrey-type updating scenario. While $x = 1/3$ as in figure 3.1, the updated $x' = 1/2$. These are again the numbers of example 11. There appear to be two plausible ways to update (y, z). One way, which we will call Jeffrey conditioning, follows the line to the origin of the coordinate system. It accords with our intuition that as $x' \to 1$, the updated (y'_1, z'_1) is continuous with standard conditioning. The other way, which in chapter 4 I will call LP conditioning, uses the shortest geometric distance from (y, z) to determine (y'_2, z'_2). This case is illustrated in three dimensions (not using statistical parameters) in figure 4.4. One of the main points of this dissertation is that the ambivalence between updating methods can be resolved by using cross-entropy rather than Euclidean distance. Jeffrey conditioning gives us the updated probability distribution which minimally diverges from the relatively prior probability distribution in terms of cross-entropy.

The affine nature of the subset guarantees that the subset is closed and convex (see Paris, 2006, 66).

> **Example 9: More Than One Half.** A coin for which I have excellent evidence that it is fair is about to be flipped. My prior probability that the coin will land tails is 0.5. A demon, whom I inductively know to be reliable about the future chances of events, tells me that on the next flip the probability for tails is greater than 0.5 (see Landes and Williamson, 2013, 3557).

If the constraint is non-affine, as in example 9, we have no such guarantees. Landes and Williamson worry about this as irrationality in an otherwise solid defence of PME (see Landes and Williamson, 2013; for more detail on affine constraints see Csiszár, 1967; Williams, 1980, 137; Howson and Franklin, 1994, 456; Uffink, 1996, 7; Paris, 2006, 66; Debbah and Müller, 2005, 1673; Williamson, 2011, 5; and Paul Penfield's lecture notes `http://www-mtl.mit.edu/Courses/6.050/2013/notes`, section 9.4). I do not address this issue here in any more depth. A comprehensive defence of PME may need to close this gap.

Note that some epistemologists defend a position which requires all affine constraints to be solved by standard conditioning (for example Jan van Campenhout and Thomas Cover, who consider PME a special case of Bayes' Theorem, see 1981; Skyrms, 1985, which van Fraassen considers "the most sensitive treatment so far" in his 1993, 313; Skyrms, 1986, 237; Howson and Franklin, 1994, 461; and Grove and Halpern, 1997, making this case for the *Judy Benjamin* problem treated in chapter 7 below).

An affine constraint makes it possible to find points in the subset that are minimally distant from the point corresponding to the prior probability distribution. Naively, we might use the Euclidean distance to determine the minimally distant point (for example in Joyce, 1998; or Leitgeb and Pettigrew, 2010a). Chapter 4 shows that such a procedure would put us at odds with a host of reasonable intuitions. In the next subsection, I provide a sketch of the information-theoretic approach.

3.2 Information Theory

3.2.1 Shannon Entropy

Information theory is rooted in the consideration of communication over noisy channels. Since most channels transmit information with errors, before Shannon the prevailing thought was that in order to reduce the probability of error, redundancy needed to be increased. No pain, no gain. Shannon's noisy-channel coding theorem proves this intuition to be false. Information can be communicated over a noisy channel at a non-zero rate with arbitrarily small error probability. The maximum rate for which this is possible is called the capacity of the channel (see MacKay, 2003, 14f).

To prove this remarkable result and calculate the capacity of a channel, Shannon needs a measure of entropy for a probability distribution. Consider the written English language and its distribution of Latin letters: the letter E occurs 9.1% of the time, whereas the letter P only occurs 1.9% of the time. There is a certain inefficiency in coding a message by giving as much space to symbols that contain less information than others. The P contains more information than an E, and in general consonants contain more information than vowels. Codes in which there are fewer of these inefficiencies have higher entropy. Entropy is therefore based on the probability distribution of the code symbols.

Let X_n be an ensemble $(x, \mathcal{A}_{X_n}, P_{X_n})$ (P_{X_n} will be abbreviated P_n in the following), where x is the outcome of a random variable, P_n is the probability distribution, and the outcome set $\mathcal{A}_{X_n}$ is an alphabet with n letters. We want a measure of entropy H_n that has the following properties (see Khinchin, 1957, for these axioms; I gleaned them from Guiaşu, 1977, 9):

(K1) For any n, the function $H_n(P_n(x_1), \ldots, P_n(x_n))$ is a continuous and symmetric function with respect to all of its arguments.

(K2) For every n, we have $H_{n+1}(P_n(x_1), \ldots, P_n(x_n), 0) = H_n(P_n(x_1), \ldots, P_n(x_n))$.

(K3) For every n, we have the inequality $H_n(P_n(x_1), \ldots, P_n(x_n)) \leq H_n(1/n, \ldots, 1/n)$.

(K4) If the conditions in (3.11) hold for π_{kl}, the consequent in (3.12) is true.

For K4, the conditions are

$$\pi_{kl} \geq 0, \sum_{k=1}^{n} \sum_{l=1}^{m_k} \pi_{kl} = 1, P_n(x_k) = \sum_{l=1}^{m_k} \pi_{kl}, M = \sum_{k=1}^{n} m_k \tag{3.11}$$

and the consequent is

$$\begin{aligned} &H_{nM}(\pi_{11}, \ldots, \pi_{nm_n}) = \\ &H_n(P_n(x_1), \ldots, P_n(x_n)) + \sum_{k=1}^{n} P_n(x_k) H_{m_k}\left(\frac{\pi_{k1}}{P_n(x_k)}, \ldots, \frac{\pi_{km_k}}{P_n(x_k)}\right). \end{aligned} \tag{3.12}$$

K1–3 are self-explanatory. K4 requires that the entropy of a product code is obtained by adding the entropy of one component to the entropy of the other component conditioned by the first one. Given Khinchin's axioms, theorem 1.1 in Silviu Guiaşu's book proves that

$$H_n(P_n(x_1), \ldots, P_n(x_n)) = -c \sum_{k=1}^{n} P_n(x_k) \log P_n(x_n) \tag{3.13}$$

Leaving off the subscript n (which is predetermined by the cardinality of the alphabet) and taking $c = 1$,

$$H(P(x_1), \ldots, P(x_n)) = -\sum_{k=1}^{n} P(x_k) \log P(x_n) \tag{3.14}$$

is called the Shannon entropy (see Shannon, 1948).

3.2.2 Kullback-Leibler Divergence

I could use the Shannon entropy as a first pass to isolate a probability distribution from a candidate set determined by an affine constraint.

> ORIGINAL PME A rational agent accepts for her partial beliefs a probability distribution that satisfies evidential constraints and has maximal Shannon entropy.

The problem with ORIGINAL PME is that it does not take into account a relatively prior probability distribution and is therefore inconsistent with the former of the two basic commitments of Bayesians: the partial beliefs of a rational agent (i) depend on a prior probability and (ii) are obtained by standard conditioning if applicable. Like the likelihood principle, ORIGINAL PME does not make reference to a prior probability and is therefore unsuitable from a Bayesian point of view.

In order to fix this, we need the concept of cross-entropy. Whereas maximizing the Shannon entropy gives us the most efficient code given a set of constraints (if they are affine constraints, this code will be uniquely distributed), minimizing cross-entropy gives us the code with respect to which the relatively prior probability distribution is most efficient. Cross-entropy measures how much information is lost when a message is coded using a code optimized for the relatively prior probability distribution rather than the 'true' posterior distribution. When I pick a posterior probability distribution according to the principle of minimum cross-entropy or *Infomin*, I pick the probability distribution that accords with my constraints and makes the relatively prior distribution look as efficient as possible. This is analogous to Bayesian conditionalization, where I choose a posterior probability distribution which strikes a balance between respecting the prior and respecting the new evidence.

To implement *Infomin*, I need a measure for cross-entropy. Again, as in K1–4, I first formulate a set of desiderata for this function that maps two probability distributions P_n and Q_n to the set of real numbers. A caveat here: the cross-entropy is conceptualized as a divergence of P_n from Q_n so that in the logic of updating Q_n is the relatively prior probability distribution and P_n is the posterior. I shall use the generic label $D(P_n, Q_n)$ for this measure in the desiderata. I shall also sometimes

write $D(P_n(x_1), \ldots, P_n(x_n); Q_n(x_1), \ldots, Q_n(x_n))$ for $D(P_n, Q_n)$. This time, I gleaned the desiderata from Paris, 2006, 121; the original proof is in Kullback and Leibler, 1951.

(KL1) For each n, $D(P_n, Q_n)$ is continuous, provided that $Q(x_i) \neq 0$ if $P(x_i) \neq 0$.

(KL2) For $0 < n \leq n_0$, $D(1/n, \ldots, 1/n, 0, \ldots, 0; 1/n_0, \ldots, 1/n_0))$ is strictly decreasing in n and increasing in n_0 and is 0 if $n = n_0$.

(KL3) Provided that $Q(x_i) \neq 0$ if $P(x_i) \neq 0$, and given a permutation σ of $(1, \ldots, n)$, the consequent (3.15) holds.

(KL4) If the conditions in (3.11) hold for π_{kl} and analogous conditions hold for ϱ_{kl} (replace P by Q), then the consequent in (3.16) is true, provided that $\varrho_{kl} \neq 0$ if $\pi_{kl} \neq 0$.

For KL3, the consequent is

$$\begin{aligned} &D(P_n(x_1), \ldots, P_n(x_n); Q_n(x_1), \ldots, Q_n(x_n)) = \\ &D(P_n(x_{\sigma(1)}), \ldots, P_n(x_{\sigma(n)}); Q_n(x_{\sigma(1)}), \ldots, Q_n(x_{\sigma(n)})) \end{aligned} \tag{3.15}$$

For KL4, the consequent is

$$\begin{aligned} &D(\pi_{11}, \ldots, \pi_{nm_n}; \varrho_{11}, \ldots, \varrho_{nm_n}) = \\ &D(P_n(x_1), \ldots, P_n(x_n); Q_n(x_1), \ldots, Q_n(x_n)) + \\ &\sum_{k=1}^{n} P_n(x_k) D\left(\frac{\pi_{k1}}{P_n(x_k)}, \ldots, \frac{\pi_{km_k}}{P_n(x_k)}; \frac{\varrho_{k1}}{P_n(x_k)}, \ldots, \frac{\varrho_{km_k}}{P_n(x_k)}\right). \end{aligned} \tag{3.16}$$

These desiderata give us the solution

$$D(P_n, Q_n) = c \sum_{i=1}^{n} P_n(x_i) \log \frac{P_n(x_i)}{Q_n(x_i)} \tag{3.17}$$

which is fulfilled by the Kullback-Leibler divergence D_{KL}, again setting $c = 1$,

$$D_{\text{KL}}(P_n, Q_n) = \sum_{i=1}^{n} P_n(x_i) \log \frac{P_n(x_i)}{Q_n(x_i)}. \tag{3.18}$$

In section 2.2, I have articulated a version of PME to be added to classical Bayesian commitments. PME seeks to be sensitive to the intuition that we ought not to gain information where the additional information is not warranted by the evidence. Some want to drive a wedge between the synchronic rule to keep the entropy maximal (PME) and the diachronic rule to keep the cross-entropy minimal (*Infomin*) (for this objection see Walley, 1991, 270f). Here is a brief excursion to dispel this worry.

> **Example 10: Piecemeal Learning.** Consider a bag with blue, red, and green tokens. You know that (C') at least 50% of the tokens are blue. Then you learn that (C'') at most 20% of the tokens are red.

The synchronic norm, on the one hand, ignores the diachronic dimension and prescribes the probability distribution which has the maximum entropy and obeys both (C') and (C''). The three-dimensional vector containing the probabilities for blue, red, and green is $(\frac{1}{2}, \frac{1}{5}, \frac{3}{10})$. The diachronic norm, on the other hand, processes (C') and (C'') sequentially, taking in its second step $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ as its prior probability distribution and then diachronically updating to $(\frac{8}{15}, \frac{1}{5}, \frac{4}{15})$.

The information provided in a problem calling for the synchronic norm and the information provided in a problem calling for the diachronic norm is different, as temporal relations and their implications for dependence between variables clearly matter. In example 10, we might have relevantly received information (C'') before (C') ('before' may be understood logically rather than temporally) so that *Infomin* updates in its last step $(\frac{2}{5}, \frac{1}{5}, \frac{2}{5})$ to $(\frac{1}{2}, \frac{1}{6}, \frac{1}{3})$. Even if (C') and (C'') are received in a definite order, the problem may be phrased in a way that indicates independence between the two constraints. In this case, the synchronic norm is the appropriate norm to use. *Infomin* does not assume such independence and therefore processes the two pieces of information separately. Disagreement arises when observations

are interpreted differently, not because PME and *Infomin* are inconsistent with each other.

The fault line for commutativity is actually located between standard conditioning (which is always commutative) and Jeffrey conditioning (which is sometimes non-commutative). Huttegger explains:

> This has caused some concerns as to whether probability kinematics is a rational way to update beliefs (Döring, 1999; Lange, 2000; and Kelly, 2008). I agree with Wagner and Joyce that these concerns are misguided (see Wagner, 2002; and Joyce, 2009). As Joyce points out, probability kinematics is non-commutative exactly when it should be; namely, when belief revision destroys information obtained in previous updates. (Huttegger, 2015, 628.)

In the following, I will assume that PME and *Infomin* are compatible and part of the toolkit at the disposal of PME.

The first question to ask is if PME is compatible with the Bayesian commitment to standard conditioning. The answer is yes. Better yet, PME also agrees with Jeffrey's extension to standard conditioning, now commonly called Jeffrey conditioning.

A proof that PME generalizes standard conditioning is in Williams, 1980. A proof that PME generalizes Jeffrey conditioning is in Jeffrey, 1965. I will give my own simple proofs here that are more in keeping with the notation in the body of the dissertation. An interested reader can also apply these proofs to show that PME generalizes Wagner conditioning, but not without simplifications that compromise mathematical rigour. The more rigorous proof for the generalization of Wagner conditioning is in section 5.4.

I assume finite (and therefore discrete) probability distributions. For countable and continuous probability distributions, the reasoning is largely analogous (for an introduction to continuous entropy see Guiaşu, 1977, 16ff; for an example of how to do a proof of this section for continuous probability densities see Caticha and Giffin, 2006, and Jaynes, 1978; for a proof that the stationary points of the Lagrange function are indeed the desired extrema see Zubarev et al., 1974, 55, and Cover and Thomas, 2006, 410; for the pioneer of the method applied in this section see Jaynes,

1978, 241ff). Before I address standard conditioning in subsection 3.2.4 and Jeffrey conditioning in subsection 3.2.5, I will present Jaynes' method in the next subsection.

3.2.3 Constraint Rule for Cross-Entropy

This subsection provides a general solution for probability kinematics using affine constraints, including standard conditioning and Jeffrey conditioning. Jaynes provides such a solution for the maximum entropy problem, best exemplified by the *Brandeis Dice* problem (see Jaynes, 1989, 243, and example 8). Jaynes' solution will not serve us because I have restricted myself to updating from relative priors rather than ignorance priors. Rather than maximizing the Shannon entropy, I will give the prescription for minimizing the Kullback-Leibler divergence. Jaynes' constraint rule can still be recovered by the case when the relative priors are equiprobabilities.

The rule is general and therefore abstract. I will make it more concrete by showing how it applies to the *Judy Benjamin* problem, which is described in more detail in section 7.1, and the story is provided in example 38. In brief, Judy Benjamin's relatively prior probabilities are $P(A_1) = 0.25, P(A_2) = 0.25, P(A_3) = 0.5$. For her posterior probabilities, the affine constraint is that $P'(A_2) = 3 \cdot P'(A_1)$. In this subsection, I will abbreviate $P(A_i) = p_i$ and $P'(A_i) = q_i$. In terms of the model introduced in section 3.1, $k = 1, b_1 = 1$ and $C = (4, 0, 1)^\intercal$ (this is not the only model; $k = 1, b_1 = 0$ and $C = (-3, 1, 0)^\intercal$ works as well). Thus, using the *Brandeis Dice* problem as a template, the *Judy Benjamin* problem is equivalent to a dice problem where the relative prior probabilities for a three-sided dice are $(0.25, 0.25, 0.5)$, the sides have 4 dots, 0 dots, and 1 dot respectively, and the average number of spots up in a long-running experiment is not 1.5 but 1 instead.

Let q_i be a posterior probability distribution on a finite space that fulfills the constraint

$$\sum_{i=1}^{n} c_i q_i = b. \tag{3.19}$$

Because q_i is a probability distribution it fulfills

$$\sum_{i=1}^{n} q_i = 1 \tag{3.20}$$

We want to minimize the Kullback-Leibler divergence from P to P', given the constraints (3.19) and (3.20),

$$\sum_{i=1}^{n} q_i \log \frac{q_i}{p_i} \tag{3.21}$$

We use Lagrange multipliers to define

$$J(q) = \sum_{i=1}^{n} q_i \log \frac{q_i}{p_i} + \lambda_0 \sum_{i=1}^{n} q_i + \lambda_1 \sum_{i=1}^{n} c_i q_i \tag{3.22}$$

and differentiate it with respect to q_i

$$\frac{\partial J}{\partial q_i} = \log q_i - \log p_i + 1 + \lambda_0 + \lambda_1 c_i \tag{3.23}$$

Set (3.23) to 0 to find the necessary condition to maximize (3.21).

$$q_i = e^{\log p_i - \lambda_0 - 1 - \lambda_1 c_i} \tag{3.24}$$

This is a combination of the Boltzmann distribution and the weight of the prior probabilities supplied by $\log p_i$. We still need to do two things: (a) show that the Kullback-Leibler divergence is minimal, and (b) show how to find λ_0 and λ_1. (a) is shown in Theorem 12.1.1 in Cover and Thomas (2006) and there is no reason to copy it here.

For (b), let

$$\beta = -\lambda_1 \tag{3.25}$$

$$Z(\beta) = \sum_{i=1}^{n} p_i e^{\beta c_i} \tag{3.26}$$

$$\lambda_0 = \log(Z(\beta)) - 1 \tag{3.27}$$

To find λ_0 and λ_1, introduce the constraint

$$\frac{\partial}{\partial \beta} \log(Z(\beta)) = b. \tag{3.28}$$

For the *Judy Benjamin* problem, substituting $x = e^{\beta}$ and filling in the known variables yields the equation

$$x^4 = \frac{4}{3} \tag{3.29}$$

with $\beta = \log x$. Use (3.25) and (3.27) to calculate $\lambda_1 = 0.27465$ and $\lambda_0 = 1.3379$. Now use (3.24) to calculate the posterior probabilities $q_1 = 0.117$, $q_2 = 0.351, q_3 = 0.533$. This result agrees with the direct method used in section 7.2 and the result obtained in (7.3).

Although it is good to know that all is well numerically, I am interested in a general proof that this choice of λ_0 and λ_1 will give me the probability distribution minimizing the cross-entropy. That q so defined minimizes the entropy is shown in (a). I need to make sure, however, that with this choice of λ_0 and λ_1 the constraints

(3.19) and (3.20) are also fulfilled.

First,

$$\sum_{i=1}^{n} q_i = \sum_{i=1}^{n} e^{\log p_i - \lambda_0 - 1 - \lambda_1 c_i} = e^{-\lambda_0 - 1} \sum_{i=1}^{n} e^{-\lambda_1 c_i} = e^{-\log(Z(\beta))} Z(\beta) = 1 \tag{3.30}$$

Then, using

$$b \sum_{i=1}^{n} p_i e^{\beta c_i} = \sum_{i=1}^{n} p_i c_i e^{\beta c_i} \tag{3.31}$$

on account of (3.28),

$$\sum_{i=1}^{n} c_i q_i = \sum_{i=1}^{n} c_i e^{\log p_i - \lambda_0 - 1 - \lambda_1 c_1} = \frac{1}{Z(\beta)} \sum_{i=1}^{n} c_i p_i e^{\beta c_1} = \frac{1}{Z(\beta)} b \sum_{i=1}^{n} p_i e^{\beta c_i} = b. \tag{3.32}$$

A proof with similar intentions as this subsection is in Paul Penfield's lecture notes `http://www-mtl.mit.edu/Courses/6.050/2013/notes`, subsection 9.6.3. Exercise 12.2. (see Cover and Thomas, 2006, 409) is nicely suggestive of the combination between weight of prior probabilities and Jaynes' constraint manifest in (3.24).

3.2.4 Standard Conditioning

Let y_i (all $y_i \neq 0$) be a finite relatively prior probability distribution summing to 1, $i \in I$. Let $\hat{y}_i$ be the posterior probability distribution derived from standard conditioning with $\hat{y}_i = 0$ for all $i \in I'$ and $\hat{y}_i \neq 0$ for all $i \in I''$, $I' \cup I'' = I$. I' and I'' specify the standard event observation. Standard conditioning requires that

$$\hat{y}_i = \frac{y_i}{\sum_{k \in I''} y_k}. \tag{3.33}$$

To solve this problem using PME, we want to minimize the cross-entropy with the constraint that the non-zero $\hat{y}_i$ sum to 1. The Lagrange function is (writing in vector form $\hat{y} = (\hat{y}_i)_{i \in I''}$)

$$\Lambda(\hat{y}, \lambda) = \sum_{i \in I''} \hat{y}_i \ln \frac{\hat{y}_i}{y_i} + \lambda \left(1 - \sum_{i \in I''} \hat{y}_i \right). \tag{3.34}$$

Differentiating the Lagrange function with respect to $\hat{y}_i$ and setting the result to zero gives us

$$\hat{y}_i = y_i e^{\lambda - 1} \tag{3.35}$$

with λ normalized to

$$\lambda = 1 - \ln \sum_{i \in I''} y_i. \tag{3.36}$$

(3.33) follows immediately. PME generalizes standard conditioning.

3.2.5 Jeffrey Conditioning

For this subsection, I am using a special notation that I develop in section 5.4. Even though thematically this subsection belongs here, it depends on understanding the notation introduced in chapter 5. The reader may want to skip this subsection and return to it later. The conclusion stands that PME generalizes Jeffrey conditioning.

Let $\theta_i, i = 1, \ldots, n$ and $\omega_j, j = 1, \ldots, m$ be finite partitions of the event space

with the joint prior probability matrix (y_{ij}) (all $y_{ij} \neq 0$). Let κ be defined as in section 5.4, with (5.11) true (noting that in section 5.4, (5.11) is no longer required). Let P be the relatively prior probability distribution and $\hat{P}$ the posterior probability distribution.

Let $\hat{y}_{ij}$ be the posterior probability distribution derived from Jeffrey conditioning with

$$\sum_{i=1}^{n} \hat{y}_{ij} = \hat{P}(\omega_j) \text{ for all } j = 1, \ldots, m \tag{3.37}$$

Jeffrey conditioning requires that for all $i = 1, \ldots, n$

$$\hat{P}(\theta_i) = \sum_{j=1}^{m} P(\theta_i|\omega_j)\hat{P}(\omega_j) = \sum_{j=1}^{m} \frac{y_{ij}}{P(\omega_j)}\hat{P}(\omega_j) \tag{3.38}$$

Using PME to get the posterior distribution $(\hat{y}_{ij})$, the Lagrange function is (writing in vector form $\hat{y} = (x_{11}, \ldots, x_{n1}, \ldots, x_{nm})^\top$ and $\lambda = (\lambda_1, \ldots, \lambda_m)^\top$)

$$\Lambda(\hat{y}, \lambda) = \sum_{i=1}^{n}\sum_{j=1}^{m} \hat{y}_{ij} \ln \frac{\hat{y}_{ij}}{y_{ij}} + \sum_{j=1}^{m} \lambda_j \left(\hat{P}(\omega_j) - \sum_{i=1}^{n} \hat{y}_{ij} \right). \tag{3.39}$$

Consequently,

$$\hat{y}_{ij} = y_{ij} e^{\lambda_j - 1} \tag{3.40}$$

with the Lagrangian parameters λ_j normalized by

$$\sum_{i=1}^{n} y_{ij} e^{\lambda_j - 1} = \hat{P}(\omega_j) \tag{3.41}$$

(3.38) follows immediately. PME generalizes Jeffrey conditioning.

3.3 Pragmatic, Axiomatic, and Epistemic Justification

3.3.1 Three Approaches

There are various ways in which an explanatory theory of a rational agent's partial beliefs can be justified. The rational agent is a relevant idealization, relevant in the sense that there may be a different explanatory theory of an actual effective human reasoner's partial beliefs. The rational agent models an inductive logic, not dissimilar to the inference relationships that model deductive logic and may also be relevant idealizations of effective human deductive reasoning.

In this dissertation, I am assuming that a rational agent already has quantitative partial beliefs (also called credences) about certain propositions and that she reasons as a Bayesian, i.e. her credences are probabilistic (obey Kolmogorov's axioms) and she updates them using standard conditioning whenever standard conditioning is applicable. This assumption is controversial. There are a few places in this dissertation where I provide some justification for them, but the focus is the next step once the Bayesian elements are already in place.

In later chapters, I shall defend the thesis that rational agents (i) have sharp credences and (ii) update them in accordance with norms based on and in agreement with information theory. Neither sharp credences nor information-based update is a necessary feature of recent Bayesian epistemology. On the contrary, most Bayesians now appear to prefer indeterminate credal states and an eclectic approach to updating. This section sets the stage by examining what a defence of my claims needs to do in order to be successful.

Partial belief epistemologists have varying preferences about what justifies their theories. For ease of exposition, I am condensing them to three major approaches: the psychological, the formal, and the philosophical. My hope is that a defence of

(i) and (ii) succeeds on all three levels. Since all three approaches have explanatory virtue, one would also hope that they converge and settle on compatible solutions to the question of which partial beliefs rational agents entertain. (Jürgen Landes and Jon Williamson similarly show how three norms converge in justifying PME: a probability norm, a calibration norm, and an equivocation norm, see Landes and Williamson, 2013.) Information theory and its virtues across all three levels support this convergence.

I now turn to a more detailed description of what I mean by the psychological, the formal, and the philosophical approaches, which I cash out as providing pragmatic, axiomatic, and epistemic justification for the partial beliefs of a rational agent.

3.3.2 The Psychological Approach

The psychological approach justifies partial beliefs by highlighting their pragmatic virtues. Frank Ramsey and Bruno de Finetti were the first to substantiate this approach (see Ramsey, 1926; and de Finetti 1931). A system of partial beliefs corresponds to a willingness to make decisions in the light of these beliefs, most helpfully to purchase a bet on a proposition p costing $\$x$ with a return of \$1 if the proposition turns out to be true and no return if the proposition turns out to be false.

Both synchronic and diachronic Dutch-book arguments demonstrate the constraints that rational agents have with respect to their partial beliefs lest they fall prey to arbitrage opportunities. The most vocal critics of the synchronic argument are Colin Howson and Peter Urbach (loc. cit., but see also Earman, 1992, 38f). The diachronic arguments, however, are significantly more controversial and less supported in the literature than the synchronic arguments originally formulated by de Finetti (for the diachronic argument supporting standard conditioning see Armendt, 1980; and Lewis, 2010; for criticism see Howson and Franklin, 1994, 458). Jon Williamson, interestingly, thinks that the diachronic argument fails because it supports Bayesian conditionalization *against* PME, with which it is incompatible and to which it must cede (see Williamson, 2011). The idea that Bayesian conditionalization and PME conflict can be found in other places as well (see Seidenfeld, 1979,

432f; Shimony, 1985; van Fraassen, 1993, 288ff; Uffink, 1995, 14; Howson and Urbach, 2006, 278; and Neapolitan and Jiang, 2014, 4012), but the usual interpretation is that the conflict undermines PME. This dissertation defends the claim that they do not conflict at all.

3.3.3 The Formal Approach

The formal approach justifies partial beliefs by providing an axiomatic system which coheres with intuitions we have about partial beliefs and limits the partial belief functions fulfilling the formal constraints. The pioneer of this method is Richard Cox in his landmark article "Probability, Frequency and Reasonable Expectation." The idea is simple and can be surprisingly powerful: formalize reasonable expectations and apply them to yield unique solutions, so that for example only partial beliefs fulfilling Bayesian requirements also fulfill the reasonable expectations. The idea is controversial, especially when it leads to unique solutions.

It is often possible to weaken the assumptions such that families of solutions, rather than unique solutions, fulfill the axioms. This is the case for Carnap's continuum of inductive methods (see Carnap, 1952) and the Rényi entropy, which generalizes Shannon's entropy and yields families of credence functions fulfilling maximum entropy requirements (see Uffink, 1995, who criticizes and weakens the assumptions of Shore and Johnson, 1980, a paradigmatic example for the formal approach supporting PME; and Huisman, 2014, who uses this idea to solve the *Judy Benjamin* problem in chapter 7 with indeterminate credal states).

There is a sense in which both the psychological and the philosophical approach are formal approaches as well, starting with a set of axioms and identifying the credence functions that fulfill them. I am setting the formal approach aside when it appears that the uniqueness result is kept in mind as a trajectory in formulating the axioms. Often, there is resistance to such a manipulation of the initial conditions in order to achieve the final result of uniqueness (see the above comment on weakening of the axioms). The formal approach, however, considers uniqueness itself a desired outcome with justificatory force. There is something at least interesting about the

set of axioms that gives us a unique solution, even though each axiom still needs to pass the rigorous test of plausibility compared to its alternatives.

For Ramsey and de Finetti, the salient axiom is invulnerability to arbitrage opportunities and does decidedly not yield uniqueness. Subjective probabilities can range wildly between rational agents, as long as they adhere to the laws of probability. Bas van Fraassen argues that calibration requirements mandate the use of standard conditioning once a relatively prior probability distribution is in place (see van Fraassen, 1989). The calibration requirement is both psychological and formal in the sense that it is based on pragmatic considerations about an agent's psychology and behaviour, especially her need to have beliefs and states of the world calibrated, but also on a formal apparatus that gives us a unique result for conditioning on new information. Joyce provides the philosophical counterpart (see Joyce, 1998), where the salient axiom is gradational accuracy (which is philosophical rather than psychological in its nature) and again standard conditioning follows as the uniquely mandated updating method fulfilling the axiom.

In a Jeffrey-type updating scenario, where standard conditioning cannot be applied, there are similar attempts to isolate a unique updating method. This method is usually Jeffrey conditioning, for which there are dynamic coherence arguments (see Armendt, 1980; Goldstein, 1983; and Skyrms, 1986), met with critical resistance in the literature (see Levi, 1987; Christensen, 1999; Talbott, 1991; Maher, 1992; and Howson and Urbach, 2006). Remarkably, Joyce's norm of gradational accuracy mandates a different unique updating method in Jeffrey-type updating scenarios, LP conditioning (see Leitgeb and Pettigrew, 2010b). I will look at this anomaly in great detail in chapter 4.

Sometimes neither standard conditioning nor Jeffrey conditioning can be applied, even though the constraint is affine. In this case, we can develop a justification of PME using the axiomatic approach (see Shore and Johnson, 1980; Tikochinsky et al., 1984; and Skilling, 1988). Criticism of PME focuses either on the strength of the assumptions (and weakening them, as in Uffink, 1996) or on deductive implications of the axioms that are inconsistent with epistemic intuitions we have about particular cases (the paradigm case is the *Judy Benjamin* problem, see van Fraassen, 1981).

The following quotes illustrate the skepticism entertained by many about PME as a general updating procedure for affine constraints:

> There has been considerable interest recently in maximum entropy methods, especially in the philosophical literature. Example 5.1 suggests that any claims to the effect that maximum-entropy revision is the only correct route to probability revision should be viewed with considerable caution because of its strong dependence on the measure of closeness being used. (Diaconis and Zabell, 1982, 829.)

> It will come as no surprise to those who have studied the relation of MAXENT to conditionalization in a larger space that there are many strategies which conflict with MAXENT and yet satisfy these conditions for coherence. (Skyrms, 1986, 241.)

> None of the arguments for the PME, when regarded as a general method for generating precise probabilities, is at all compelling. (Walley, 1991, 271.)

> The fact that Jeffrey's rule coincides with MAXENT is simply a misleading fluke, put in its proper perspective by the natural generalization of Jeffrey conditionalization described in this paper. (Wagner, 1992, 255.)

> We saw that simple conditionalization is actually the rule enjoined by the principle of minimum information in such circumstances. Furthermore, not to use the rule of Bayesian conditionalization, but some other rule, like the principle of minimum information with a uniform prior and constraints in the form of expectation values, actually entails inconsistency, i.e. incoherence. (Howson and Franklin, 1994, 465.)

> Maximum entropy is also sometimes proposed as a method for solving inference problems . . . I think it is a bad idea to use maximum entropy in this way; it can give very silly answers. (MacKay, 2003, 308.)

> Maximum entropy and relative entropy have proved quite successful in a number of applications, from physics to natural-language modeling. Unfortunately, they also exhibit some counterintuitive behavior on certain applications. Although they are valuable tools, they should be used with care. (Halpern, 2003, 110.)

> [PME] essentially never gives the right results ... it is likely to give highly misleading answers. (Grünwald and Halpern, 2003, 243.)

> Given the other problematic features of conditionalization we pointed to in Chapter 3, we feel that in linking its fortunes to the principle of minimum information no real advance has been made in justifying its adoption as an independent Bayesian principle. (Howson and Urbach, 2006, 287f.)

> It is notoriously difficult to defend general procedures for directly updating credences on constraints. (Moss, 2013, 7.)

My dissertation disagrees with these assessments. The axioms which require a unique updated probability distribution that is minimally informative with respect to the prior probability distribution are defensible and do not lead to incoherence.

3.3.4 The Philosophical Approach

The philosophical approach justifies partial beliefs by highlighting epistemic virtues of partial beliefs, contrasted usually with pragmatic virtues. The idea is that there is independent appeal in having doxastic states that are as close as possible to the truth, no matter what their pragmatic consequences are. The manifesto of this approach is James Joyce's article "A Nonpragmatic Vindication of Probabilism." In full belief epistemology, epistemic virtue consists in the balance between believing as many truths and disbelieving as many falsehoods as possible. A trade-off may be involved between valuing a low error ratio and valuing a high number of beliefs (on which one can act, for example, and be better safe than sorry, even when the belief is false, see Stephens, 2001).

The point of this approach, however, is that pragmatic considerations take the back seat and give priority to a rational agent's desire to get as close as possible to reflecting the state of the world in her epistemic state. Joyce's *Norm of Gradational Accuracy*, which generalizes the full belief norm of accuracy to partial beliefs, succeeds in affording us a system of requirements with substantial epistemic implications, such as probabilism, Bayes' formula, perhaps a principle of indifference,

and other forms of conditioning (see Greaves and Wallace, 2006; and Leitgeb and Pettigrew, 2010a).

As different as the three approaches are, they intersect on a significant amount of common terrain. Rational agents are subject to norms in their partial beliefs because we have intuitions about deficient partial beliefs. It is not the laws of logic which circumscribe these norms. The disagreement is often about how far intuition can take us in this type of circumscription. If we go too far, there is a danger of counter-examples to overly narrow norms or even inconsistencies between the norms. Often, epistemologists feel that one should only go as far as necessary, with an increasingly heavy burden of proof as indeterminacies give way to determinate solutions. If we do not go far enough, weak norms licence partial beliefs that are counter-intuitive to some. Much of the debate centres around striking an intuitively plausible balance between these two impulses.

This dissertation defends an approach to norms for the partial beliefs of a rational agent that many may find constricting. In scientific practice, however, there may be an advantage to a relatively specific normative theory of partial beliefs. The scientist does not need to consult the philosopher on questions of updating procedures, for example, if the simple intuitions of information theory reliably work with their established formal methods. I will address this advantage again when I talk about the full employment theorem (see section 7.1).

There are examples from the history of partial belief norms which reflect the tension between the relatively liberal and relatively conservative approach. Carnap's continuum of inductive methods, even though as a continuum with variable lambda it pays homage to liberalism (which E.T. Jaynes subsequently criticized by fixing $\lambda = 2$, see Jaynes and Bretthorst, 2003), meets resistance primarily from those who consider Carnap's assumptions too strong and his conclusions therefore too narrow. Bayesians are constantly confronted with examples where standard conditioning is supposed to give the wrong answer (see Howson and Urbach, 2006, 81; or Williamson, 2011).

At the foundation of my work is the intuition that allowing violations of information theory's principles saddles us with "absurdities that qualify as rational" (quoting

from Salmon, 1967, 81, in a different context), although these absurdities are often everything but plain. Whether or not this intuition is defensible rests on the confirmation and disconfirmation gleaned from counter-examples, conceptual questions, and the integrity and power of formal accounts.

Chapter 4

Asymmetry and the Geometry of Reason

4.1 Contours of a Problem

In the early 1970s, the dominant models for similarity in the psychological literature were all geometric in nature. Distance measures capturing similarity and dissimilarity between concepts obeyed minimality, symmetry, and the triangle inequality. Then Amos Tversky wrote a compelling paper undermining the idea that a metric topology is the best model (see Tversky, 1977). Tversky gave both theoretical and empirical reasons why similarity between concepts fulfills neither minimality, nor symmetry, nor the triangle inequality. Geometry with its metric distance measures was in some ways not a useful model of similarity. Tversky presented an alternative set-theoretic account which accommodated intuitions that could not be reconciled with a geometry of similarity.

The aim of this chapter is to help along a similar paradigm shift when it comes to epistemic modeling of closeness or difference between subjective probability distributions. The 'geometry of reason' (a term coined by Richard Pettigrew and Hannes Leitgeb, two of its advocates) violates reasonable expectations for an acceptable model. A non-metric alternative, information theory, fulfills many of these expectations but violates others which are similarly intuitive. Instead of presenting a third alternative which coheres better with the list of expectations outlined in section 4.4, I defend the view that while the violations of the geometry of reason are irremediable, there is a promise in the wings that an advanced formal account of information theory, using the theory of differential manifolds, can explain information theory's

violations of prima facie reasonable expectations.

The geometry of reason refers to a view of epistemic utility in which the underlying topology for credence functions (which may be subjective probability distributions) on a finite number of events is a metric space. The set of non-negative credences that an agent assigns to the outcome of a die roll, for example, is isomorphic to $\mathbb{R}^6_{\geq 0}$. If the agent fulfills the requirements of probabilism, the isomorphism is to the more narrow set $\mathbb{S}^5$, the five-dimensional simplex for which

$$p_1 + p_2 + p_3 + p_4 + p_5 + p_6 = 1. \tag{4.1}$$

For the remainder of this paper I will assume probabilism and an isomorphism between probability distributions P on an outcome space Ω with $|\Omega| = n$ and points $p \in \mathbb{S}^{n-1} \subset \mathbb{R}^n$ having coordinates $p_i = P(\omega_i), i = 1, \ldots, n$ and $\omega_i \in \Omega$. Since the isomorphism is to a metric space, there is a distance relation between credence functions which can be used to formulate axioms relating credences to epistemic utility and to justify or to criticize contentious positions such as Bayesian conditionalization, the principle of indifference, other forms of conditioning, or probabilism itself (see especially works cited below by James Joyce; Pettigrew and Leitgeb; David Wallace and Hilary Greaves). For information theory, as opposed to the geometry of reason, the underlying topology for credence functions is not a metric space (see figures 4.1 and 4.2 for illustration).

Epistemic utility in Bayesian epistemology has attracted some attention in the past few years. Patrick Maher provides a compelling acceptance-based account of epistemic utility (see Maher, 1993, 182–207). Joyce, in "A Nonpragmatic Vindication of Probabilism," defends probabilism supported by partial-belief-based epistemic utility rather than the pragmatic utility common in Dutch-book style arguments (see Joyce, 1998). For Joyce, norms of gradational accuracy characterize the epistemic utility approach to partial beliefs, analogous to norms of truth for full beliefs.

Wallace and Greaves investigate epistemic utility functions along 'stability' lines and conclude that for everywhere stable utility functions standard conditioning is

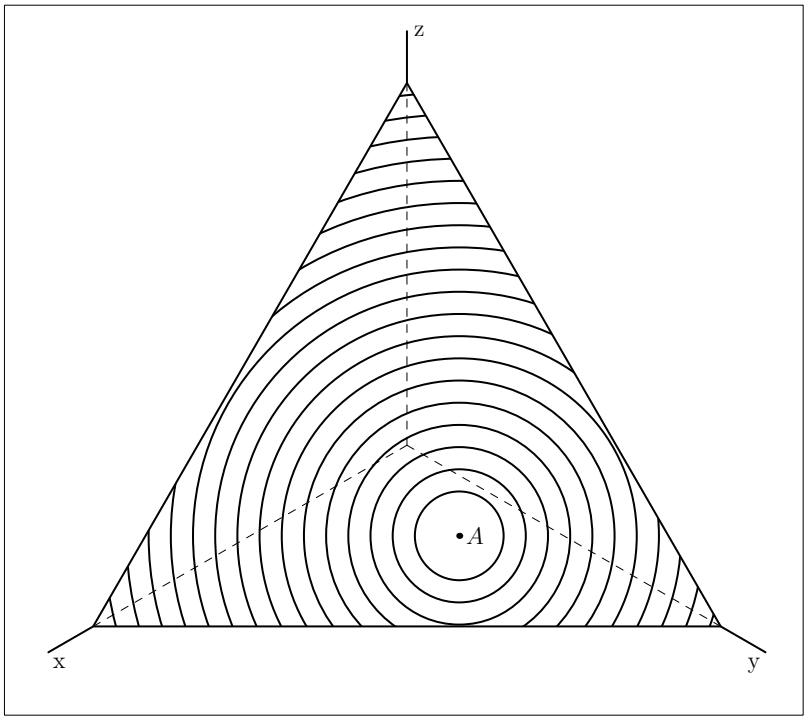

Figure 4.1: The simplex $\mathbb{S}^2$ in three-dimensional space $\mathbb{R}^3$ with contour lines corresponding to the geometry of reason around point A in equation (4.12). Points on the same contour line are equidistant from A with respect to the Euclidean metric. Compare the contour lines here to figure 4.2. Note that this diagram and all the following diagrams are frontal views of the simplex.

optimal, while only somewhere stable utility functions create problems for maximizing expected epistemic utility norms (see Greaves and Wallace, 2006; and Pettigrew, 2013). Richard Pettigrew and Hannes Leitgeb have published arguments that under certain assumptions probabilism and standard conditioning (which together give epistemology a distinct Bayesian flavour) minimize inaccuracy, thereby providing maximal epistemic utility (see Leitgeb and Pettigrew, 2010a and 2010b).

Leitgeb and Pettigrew show, given the geometry of reason and other axioms inspired by Joyce (for example normality and dominance), that in order to avoid epistemic dilemmas we must commit ourselves to a Brier score measure of inaccuracy and subsequently to probabilism and standard conditioning. The Brier score is the

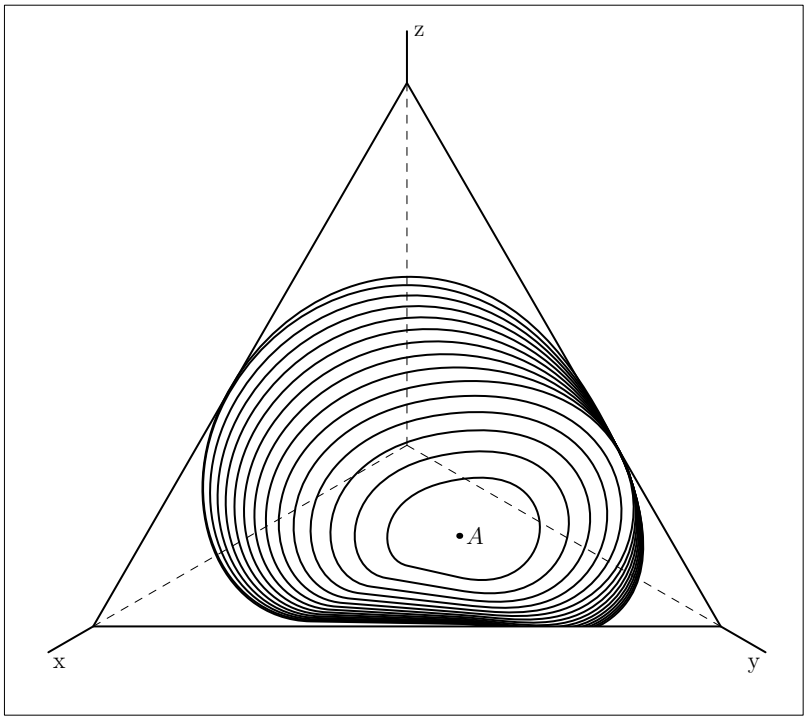

Figure 4.2: The simplex $\mathbb{S}^2$ with contour lines corresponding to information theory around point A in equation (4.12). Points on the same contour line are equidistant from A with respect to the Kullback-Leibler divergence. The contrast to figure 4.1 will become clear in much more detail in the body of the chapter. Note that the contour lines of the geometry of reason are insensitive to the boundaries of the simplex, while the contour lines of information theory reflect them. One of the main arguments in this chapter is that information theory respects epistemic intuitions we have about asymmetry: proximity to extreme beliefs with very high or very low probability influences the topology that is at the basis of updating.

mean squared error of a probabilistic forecast. For example, if we look at 100 days for which the forecast was 30% rain and the incidence of rain was 32 days, then the Brier score is

$$\frac{1}{N}\sum_{t=1}^{N}(f_t - o_t) = \frac{1}{100}\left(32\cdot(0.3-1) + 68\cdot(0.3-0)\right) = 0.218 \tag{4.2}$$

0 is a perfect match between forecast and reality in the sense that the forecaster anticipates every instance of rain with a 100% forecast and every instance of no rain with a 0% forecast.

Jeffrey conditioning (also called probability kinematics) is widely considered to be a commonsense extension of standard conditioning. On Leitgeb and Pettigrew's account, using the Brier score, it fails to provide maximal epistemic utility. Another type of conditioning, which we will call LP conditioning, takes the place of Jeffrey conditioning. The failure of Jeffrey conditioning to minimize inaccuracy on the basis of the geometry of reason casts, by reductio, doubt on the geometry of reason.

I will show that LP conditioning, which the geometry of reason entails, fails commonsense expectations that are reasonable to have for the kind of updating scenario that LP conditioning addresses. To relate probability distributions to each other geometrically, using the isomorphism between the set of probability distributions on a finite event space W with $|W| = n$ and the $n-1$-dimensional simplex $\mathbb{S}^{n-1} \subset \mathbb{R}^n$, is initially an arbitrary move. Leitgeb and Pettigrew do little to substantiate a link between the geometry of reason and epistemic utility on a conceptual level. It is the formal success of the model that makes the geometry of reason attractive, but the failure of LP conditioning to meet basic expectations undermines this success.

The question then remains whether we have a plausible candidate to supplant the geometry of reason. The answer is yes: information theory provides us with a measure of closeness between probability distributions on a finite event space that has more conceptual appeal than the geometry of reason, especially with respect to epistemic utility—it is intuitively correct to relate coming-to-knowledge to exchange of information. More persuasive than intuition, however, is the fact that information theory supports both standard conditioning (see Williams, 1980) and the extension of standard conditioning to Jeffrey conditioning (see subsection 3.2.5), an extension which is on the one hand intuitive (see Wagner, 2002) and on the other hand formally continuous with the standard conditioning which Leitgeb and Pettigrew have worked so hard to vindicate nonpragmatically. LP conditioning is not continuous with standard conditioning, which is reflected in one of the expectations that LP conditioning fails to meet.

4.2 Epistemic Utility and the Geometry of Reason

4.2.1 Epistemic Utility for Partial Beliefs

There is more epistemic virtue for an agent in believing a truth rather than not believing it and in not believing a falsehood rather than believing it. Accuracy in full belief epistemology can be measured by counting four sets, believed truths and falsehoods as well as unbelieved truths and falsehoods, and somehow relating them to each other such that epistemic virtue is rewarded and epistemic vice penalized. Accuracy in partial belief epistemology must take a different shape since as a 'guess' all partial non-full beliefs are off the mark so that they need to be appreciated as 'estimates' instead. Richard Jeffrey distinguishes between guesses and estimates: a guess fails unless it is on target, whereas an estimate succeeds depending on how close it is to the target.

The gradational accuracy needed for partial belief epistemology is reminiscent of verisimilitude and its associated difficulties in the philosophy of science (see Popper, 1963; Gemes, 2007; and Oddie, 2013). Joyce and Leitgeb/Pettigrew propose various axioms for a measure of gradational accuracy for partial beliefs relying on the geometry of reason, i.e. the idea of geometrical distance between distributions of partial belief expressed in non-negative real numbers. In Joyce, a metric space for probability distributions is adopted without much reflection. The midpoint between two points, for example, which is freely used by Joyce, assumes symmetry between the end points. The asymmetric divergence measure that I propose as an alternative to the Euclidean distance measure has no meaningful concept of a midpoint.

Leitgeb and Pettigrew muse about alternative geometries, especially non-Euclidean ones. They suspect that these would be based on and in the end reducible to Euclidean geometry but they do not entertain the idea that they could drop the requirement of a metric topology altogether (for the use of non-Euclidean geodesics in statistical inference see Shun-ichi, 1985). Thomas Mormann explicitly warns against the assumption that the metrics for a geometry of logic is Euclidean by default: "All

too often, we rely on geometric intuitions that are determined by Euclidean prejudices. The geometry of logic, however, does not fit the standard Euclidean metrical framework" (see Mormann, 2005, 433; also Miller, 1984). Mormann concludes in his article "Geometry of Logic and Truth Approximation,"

> Logical structures come along with ready-made geometric structures that can be used for matters of truth approximation. Admittedly, these geometric structures differ from those we are accostumed [sic] with, namely, Euclidean ones. Hence, the geometry of logic is not Euclidean geometry. This result should not come as a big surprise. There is no reason to assume that the conceptual spaces we use for representing our theories and their relations have an Euclidean structure. On the contrary, this would appear to be an improbable coincidence. (Mormann, 2005, 453.)

4.2.2 Axioms for Epistemic Utility

Leitgeb and Pettigrew present the following salient axioms (see Leitgeb and Pettigrew, 2010a, 219):

> **Local Normality and Dominance**: If I is a legitimate inaccuracy measure, then there is a strictly increasing function $f : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ such that, for any $A \in W$, $w \in W$, and $x \in \mathbb{R}_0^+$,
>
> $$I(A, w, x) = f\left(|\chi_A(w) - x|\right). \tag{4.3}$$
>
> **Global Normality and Dominance**: If G is a legitimate global inaccuracy measure, there is a strictly increasing function $g : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ such that, for all worlds w and belief functions $b \in \mathrm{Bel}(W)$,
>
> $$G(w, b) = g\left(\|w - b_{\mathrm{glo}}\|\right). \tag{4.4}$$

Similarly to Joyce, these axioms are justified on the basis of geometry, but this time more explicitly so:

> Normality and Dominance [are] a consequence of taking seriously the talk of inaccuracy as 'distance' from the truth, and [they endorse] the geometrical picture provided by

> Euclidean n-space as the correct clarification of this notion. As explained in section 3.2, the assumption of this geometrical picture is one of the presuppositions of our account, and we do not have much to offer in its defense, except for stressing that we would be equally interested in studying the consequences of minimizing expected inaccuracy in a non-Euclidean framework. But without a doubt, starting with the Euclidean case is a natural thing to do.

Leitgeb and Pettigrew define two notions, local and global inaccuracy, and show that one must adopt a Brier score to measure inaccuracy in order to avoid epistemic dilemmas trying to minimize inaccuracy on both measures. To give the reader an idea what this looks like in detail and for purposes of later exposition, I want to provide some of the formal apparatus. Let W be a set of worlds and $A \subseteq W$ a proposition. Then

$$I : P(W) \times W \times \mathbb{R}_0^+ \to \mathbb{R}_0^+ \tag{4.5}$$

is a measure of local inaccuracy such that $I(A, w, x)$ measures the inaccuracy of the degree of credence x with respect to A at world w. Let $\mathrm{Bel}(W)$ be the set of all belief functions (what we have been calling distributions of partial belief). Then

$$G : W \times \mathrm{Bel}(W) \to \mathbb{R}_0^+ \tag{4.6}$$

is a measure of global inaccuracy of a belief function b at a possible world w such that $G(w, b)$ measures the inaccuracy of a belief function b at world w.

Axioms such as normality and dominance guarantee that the only legitimate measures of inaccuracy are Brier scores if one wants to avoid epistemic dilemmas where one receives conflicting advice from the local and the global measures. For local inaccuracy measures, this means that there is $\lambda \in \mathbb{R}^+$ such that

$$I(A, w, x) = \lambda \left(\chi_A(w) - x\right)^2 \tag{4.7}$$

where χ_A is the characteristic function of A. For global inaccuracy measures, this means that there is $\mu \in \mathbb{R}^+$ such that

$$G(w, b) = \mu \|w - b\|^2 \tag{4.8}$$

where w and b are represented by vectors and $\|u - v\|$ is the Euclidean distance

$$\sqrt{\sum_{i=1}^{n} (u_i - v_i)^2}. \tag{4.9}$$

We use (4.7) to define expected local inaccuracy of degree of belief x in proposition A by the lights of belief function b, with respect to local inaccuracy measure I, and over the set E of epistemically possible worlds as follows:

$$\mathrm{LExp}_b(I, A, E, x) = \sum_{w \in E} b(\{w\}) I(A, w, x) = \sum_{w \in E} b(\{w\}) \lambda \left(\chi_A(w) - x\right)^2. \tag{4.10}$$

We use (4.8) to define expected global inaccuracy of belief function b' by the lights of belief function b, with respect to global inaccuracy measure G, and over the set E of epistemically possible worlds as follows:

$$\mathrm{GExp}_b(G, E, b') = \sum_{w \in E} b(\{w\}) G(w, b') = \sum_{w \in E} b(\{w\}) \mu \|w - b\|^2. \tag{4.11}$$

To give a flavour of how attached the axioms are to the geometry of reason, here are Joyce's axioms called Weak Convexity and Symmetry, which he uses to justify

probabilism. Note that in terms of notation I for Joyce is global and related to Leitgeb and Pettigrew's G in (4.6) rather than I in (4.5):

> **Weak Convexity**: Let $m = (0.5b' + 0.5b'')$ be the midpoint of the line segment between b' and b''. If $I(b', \omega) = I(b'', \omega)$, then it will always be the case that $I(b', \omega) \geq I(m, \omega)$ with identity only if $b' = b''$.
>
> **Symmetry**: If $I(b', \omega) = I(b'', \omega)$, then for any $\lambda \in [0, 1]$ one has $I(\lambda b' + (1 - \lambda)b'', \omega) = I((1 - \lambda)b' + \lambda b''), \omega)$.

Joyce advocates for these axioms in Euclidean terms, using justifications such as "the change in belief involved in going from b' to b'' has the same direction but a doubly greater magnitude than change involved in going from b' to [the midpoint] m" (see Joyce, 1998, 596). In section 4.5.3, I will show that Weak Convexity holds, and Symmetry does not hold, in 'information geometry,' the topology generated by the Kullback-Leibler divergence. The term information geometry is due to Imre Csiszár, who considers the Kullback-Leibler divergence a non-commutative (asymmetric) analogue of squared Euclidean distance and derives several results that are intuitive information geometric counterparts of standard results in Euclidean geometry (see chapter 3 of Csiszár and Shields, 2004).

4.2.3 Expectations for Jeffrey-Type Updating Scenarios

Leitgeb and Pettigrew's work is continuous with Joyce's work, but significantly goes beyond it. Joyce wants weaker assumptions and would be leery of expected inaccuracies (4.10) and (4.11), as they might presuppose the probabilism that Joyce wants to justify. Leitgeb and Pettigrew investigate not only whether probabilism and standard conditioning follow from gradational accuracy based on the geometry of reason, but also uniform distribution (their term for the claim of objective Bayesians that there is some principle of indifference for ignorance priors) and Jeffrey conditioning. They show that uniform distribution requires additional axioms which are much less plausible than the ones on the basis of which they derive probabilism and standard

conditioning (see Leitgeb and Pettigrew, 2010b, 250f); and that Jeffrey conditioning does not fulfill Joyce's Norm of Gradational Accuracy (see Joyce, 1998, 579) and therefore violates the pursuit of epistemic virtue. Leitgeb and Pettigrew provide us with an alternative method of updating for Jeffrey-type updating scenarios, which I will call LP conditioning.

> **Example 11: Sherlock Holmes.** Sherlock Holmes attributes the following probabilities to the propositions E_i that k_i is the culprit in a crime: $P(E_1) = 1/3, P(E_2) = 1/2, P(E_3) = 1/6$, where k_1 is Mr. R., k_2 is Ms. S., and k_3 is Ms. T. Then Holmes finds some evidence which convinces him that $P'(F^*) = 1/2$, where F^* is the proposition that the culprit is male and P is relatively prior to P'. What should be Holmes' updated probability that Ms. S. is the culprit?

I will look at the recommendations of Jeffrey conditioning and LP conditioning for example 11 in the next section. For now note that LP conditioning violates all of the following plausible expectations in List One for an amujus, an 'alternative method of updating for Jeffrey-type updating scenarios.' This is List One:

- CONTINUITY An amujus ought to be continuous with standard conditioning as a limiting case.
- REGULARITY An amujus ought not to assign a posterior probability of 0 to an event which has a positive prior probability and about which the intervening evidence says nothing except that a strictly weaker event has a positive posterior probability.
- LEVINSTEIN An amujus ought not to give "extremely unattractive" results in a Levinstein scenario (see Levinstein, 2012, which not only articulates this failed expectation for LP conditioning, but also the previous two).
- INVARIANCE An amujus ought to be partition invariant.
- EXPANSIBILITY An amujus ought to be insensitive to an expansion of the event space by zero-probability events.

- CONFIRMATION An amujus ought to align with intuitions we have about degrees of confirmation.

- HORIZON An amujus ought to exhibit the horizon effect which makes probability distributions which are nearer to extreme probability distributions appear to be closer to each other than they really are.

Jeffrey conditioning and LP conditioning are both an amujus based on a concept of quantitative difference between probability distributions measured as a function on the isomorphic manifold (in our case, an $n-1$-dimensional simplex). Evidence appears in the form of a constraint on acceptable probability distributions and the closest acceptable probability to the orginal (relatively prior) probability distribution is chosen as its successor. Here is List Two, a list of reasonable expectations one may have toward this concept of quantitative difference (we call it a distance function for the geometry of reason and a divergence for information theory). Let $d(p, q)$ express this concept mathematically.

- TRIANGULARITY The concept obeys the triangle inequality. If there is an intermediate probability distribution, it will not make the difference smaller: $d(p, r) \leq d(p, q) + d(q, r)$. Buying a pair of shoes is not going to be more expensive than buying the two shoes individually.

- COLLINEAR HORIZON This expecation is just a more technical restatement of the HORIZON expectation in the previous list. If p, p', q, q' are collinear with the centre of the simplex m (whose coordinates are $m_i = 1/n$ for all i) and an arbitrary but fixed boundary point $\xi \in \partial\mathbb{S}^{n-1}$ and p, p', q, q' are all between m and ξ with $\|p' - p\| = \|q' - q\|$ where p is strictly closest to m, then $|d(p, p')| < |d(q, q')|$. For an illustration of this expectation see figure 4.3. The absolute value is added as a feature to accommodate degree of confirmation functions in subsection 4.4.7, which may be negative.

- TRANSITIVITY OF ASYMMETRY An ordered pair (p, q) of simplex points associated with probability distributions is asymmetrically negative, positive, or balanced, so either $d(p, q) - d(q, p) < 0$ or $d(p, q) - d(q, p) > 0$ or $d(p, q) - d(q, p) = 0$.

If (p, q) and (q, r) are asymmetrically positive, (p, r) ought not to be asymmetrically negative. Think of a bicycle route map with different locations at varying altitudes. If it takes 20 minutes to get from A to B but only 15 minutes to get from B to A then (A, B) is asymmetrically positive. If (A, B) and (B, C) are asymmetrically positive, then (A, C) ought not to be asymmetrically negative.

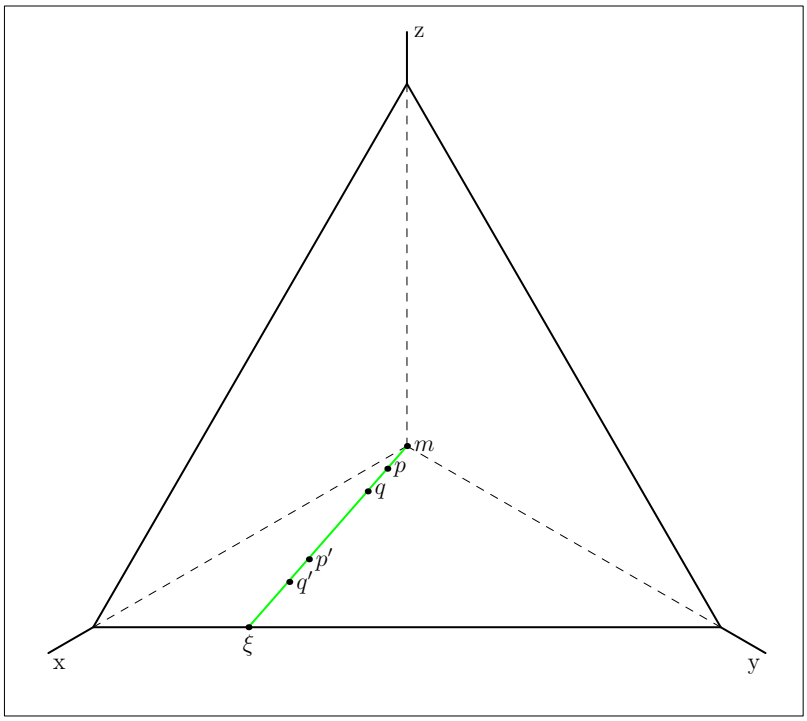

Figure 4.3: An illustrations of conditions (i)–(iii) for COLLINEAR HORIZON in List Two. p, p' and q, q' must be equidistant and collinear with m and ξ. If q, q' is more peripheral than p, p', then COLLINEAR HORIZON requires that $|d(p, p')| < |d(q, q')|$.

While the Kullback-Leibler divergence of information theory fulfills all the expectations of List One, save HORIZON, it fails all the expectations in List Two. Obversely, the Euclidean distance of the geometry of reason fulfills all the expectations of List Two, save COLLINEAR HORIZON, and fails all the expectations in List One. Information theory has its own axiomatic approach to justifying probabilism and standard

conditioning (see Shore and Johnson, 1980). Information theory provides a justification for Jeffrey conditioning and generalizes it (see subsection 3.2.5). All of these virtues stand in contrast to the violations of the expectations in List Two. The rest of this chapter fills in the details of these violations both for the geometry of reason and information theory, with the conclusion that the case for the geometry of reason is hopeless while the case for information theory is now a major challenge for future research projects.

4.3 Geometry of Reason versus Information Theory

Here is a simple example where the distance of geometry and the divergence of information theory differ. With this difference in mind, I will show how LP conditioning fails the expectations outlined in List One (see page 74). Consider the following three points in three-dimensional space:

$$a = \left(\frac{1}{3}, \frac{1}{2}, \frac{1}{6}\right) \qquad b = \left(\frac{1}{2}, \frac{3}{8}, \frac{1}{8}\right) \qquad c = \left(\frac{1}{2}, \frac{5}{12}, \frac{1}{12}\right) \tag{4.12}$$

All three are elements of the simplex $\mathbb{S}^2$: their coordinates add up to 1. Thus they represent probability distributions A, B, C over a partition of the event space into three events. The Kullback-Leibler divergence and Euclidean distance give different recommendations with respect to proximity. Note that the Kullback-Leibler divergence defined in (3.18) is always positive as a consequence of Gibbs' inequality, irrespective of dimension, (see MacKay, 2003, sections 2.6 and 2.7). The Euclidean distance $\|B - A\|$ is defined as in equation (4.9). What is remarkable about the three points in (4.12) is that

$$\|C - A\| \approx 0.204 < \|B - A\| \approx 0.212 \tag{4.13}$$

and

$$D_{\mathrm{KL}}(B, A) \approx 0.0589 < D_{\mathrm{KL}}(C, A) \approx 0.069. \tag{4.14}$$

Assuming the global inaccuracy measure GExp presented in (4.8) and $E = W$ (all possible worlds are epistemically accessible),

$$\mathrm{GExp}_A(C) \approx 0.653 < \mathrm{GExp}_A(B) \approx 0.656. \tag{4.15}$$

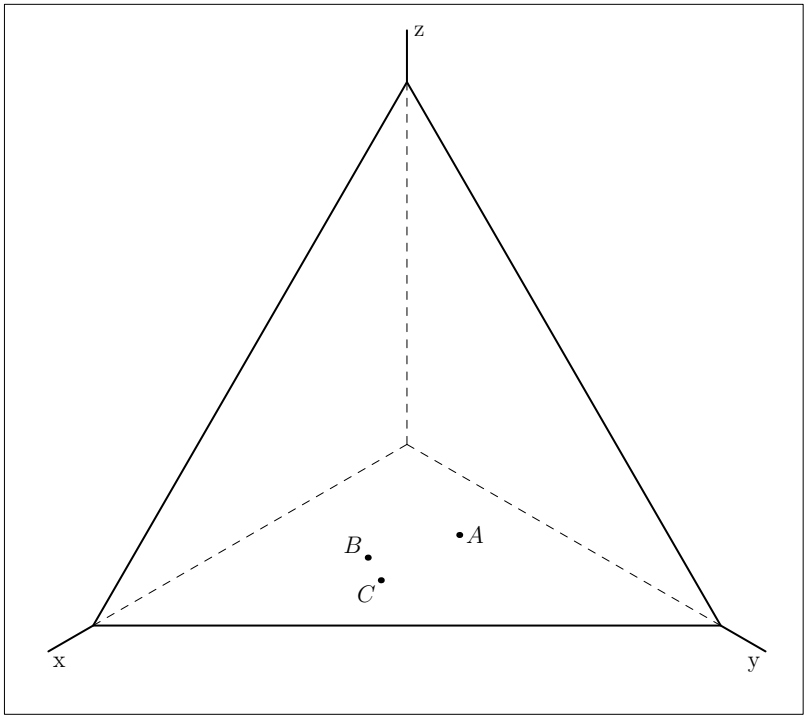

Figure 4.4: The simplex $\mathbb{S}^2$ in three-dimensional space $\mathbb{R}^3$ with points a, b, c as in equation (4.12) representing probability distributions A, B, C. Note that geometrically speaking C is closer to A than B is. Using the Kullback-Leibler divergence, however, B is closer to A than C is. The same case modeled with statistical parameters is illustrated in figures 3.1 and 3.2.

Global inaccuracy reflects the Euclidean proximity relation, not the recommendation of information theory. If A corresponds to my prior and my evidence is such that I must change the first coordinate to 1/2 (as in example 11) and nothing stronger, then information theory via the Kullback-Leibler divergence recommends the posterior corresponding to B; and the geometry of reason as expounded in Leitgeb and Pettigrew recommends the posterior corresponding to C. There are several things going on here that need some explanation.

4.3.1 Evaluating Partial Beliefs in Light of Others

We note that for Leitgeb and Pettigrew, expected global inaccuracy of b' is always evaluated by the lights of another partial belief distribution b. This may sound counterintuitive. Should we not evaluate b' by its own lights? It is part of a larger Bayesian commitment that partial belief distributions are not created ex nihilo. They can also not be evaluated for inaccuracy ex nihilo. Leitgeb and Pettigrew say very little about this, but it appears that there is a deeper problem here with the flow of diachronic updating. The classic Bayesian picture is one of moving from a relatively prior probability distribution to a posterior distribution (distinguish relatively prior probability distributions, which precede posterior probability distributions in updating, from absolutely prior probability distributions, which are ignorance priors in the sense that they are not the resulting posteriors of previous updating). This is nicely captured by standard conditioning, Bayes' formula, and updating on the basis of information theory (Jeffrey conditioning, principle of maximum entropy).

The geometry of reason and notions of accuracy based on it sit uncomfortably with this idea of flow, as the suggestion is that partial belief distributions are evaluated on their accuracy without reference to prior probability distributions—why should the accuracy or epistemic virtue of a posterior probability distribution depend on a prior probability distribution which has already been debunked by the evidence? I agree with Leitgeb and Pettigrew that there is no alternative here but to evaluate the posterior by the lights of the prior. Not doing so would saddle us with Carnap's Straight Rule, where priors are dismissed as irrelevant (see Carnap,

1952, 40ff). Yet we shall note that a justification of evaluating a belief function's accuracy by the lights of another belief function is a lot less persuasive than the way Bayesians and information theory integrate prior distributions into forming posterior distributions by virtue of an asymmetric flow of information (see also Shogenji, 2012, who makes a strong case for the influence of prior probabilities on epistemic justification).

4.3.2 LP conditioning and Jeffrey Conditioning

I want to outline how Leitgeb and Pettigrew arrive at posterior probability distributions in Jeffrey-type updating scenarios. I will call their method LP conditioning.

> **Example 12: Abstract Holmes.** Consider a possibility space $W = E_1 \cup E_2 \cup E_3$ (the E_i are sets of states which are pairwise disjoint and whose union is W) and a partition $\mathcal{F}$ of W such that $\mathcal{F} = \{F^*, F^{**}\} = \{E_1, E_2 \cup E_3\}$.

Let P be the prior probability function on W and P' the posterior. I will keep the notation informal to make this simple, not mathematically precise. Jeffrey-type updating scenarios give us new information on the posterior probabilities of partitions such as $\mathcal{F}$. In example 12, let

$$\begin{aligned} P(E_1) &= 1/3 \\ P(E_2) &= 1/2 \\ P(E_3) &= 1/6 \end{aligned} \tag{4.16}$$

and the new evidence constrain P' such that $P'(F^*) = 1/2 = P'(F^{**})$.

Jeffrey conditioning works on the following intuition, which elsewhere I have called Jeffrey's updating principle JUP (see also Wagner, 2002). The posterior probabilities conditional on the partition elements equal the prior probabilities conditional on the partition elements since we have no information in the evidence that they should have changed. Hence,

$$\begin{aligned} P'_{\text{JC}}(E_i) &= P'(E_i|F^*)P'(F^*) + P'(E_i|F^{**})P'(F^{**}) \\ &= P(E_i|F^*)P'(F^*) + P(E_i|F^{**})P'(F^{**}) \end{aligned} \tag{4.17}$$

Jeffrey conditioning is controversial (for an introduction to Jeffrey conditioning see Jeffrey, 1965; for its statistical and formal properties see Diaconis and Zabell, 1982; for a pragmatic vindication of Jeffrey conditioning see Armendt, 1980, and Skyrms, 1986; for criticism see Howson and Franklin, 1994). Information theory, however, supports Jeffrey conditioning. Leitgeb and Pettigrew show that Jeffrey conditioning does not in general pick out the minimally inaccurate posterior probability distribution. If the geometry of reason as presented in Leitgeb and Pettigrew is sound, this would constitute a powerful criticism of Jeffrey conditioning. Leitgeb and Pettigrew introduce an alternative to Jeffrey conditioning, which we have called LP conditioning. It proceeds as follows for example 12 and in general provides the minimally inaccurate posterior probability distribution in Jeffrey-type updating scenarios.

Solve the following two equations for x and y:

$$\begin{aligned} P(E_1) + x &= P'(F^*) \\ P(E_2) + y + P(E_3) + y &= P'(F^{**}) \end{aligned} \tag{4.18}$$

and then set

$$\begin{aligned} P'_{\text{LP}}(E_1) &= P(E_1) + x \\ P'_{\text{LP}}(E_2) &= P(E_2) + y \\ P'_{\text{LP}}(E_3) &= P(E_3) + y \end{aligned} \tag{4.19}$$

For the more formal and more general account see Leitgeb and Pettigrew, 2010b, 254. The results for example 12 are:

$$\begin{aligned} P'_{\text{LP}}(E_1) &= 1/2 \\ P'_{\text{LP}}(E_2) &= 5/12 \\ P'_{\text{LP}}(E_3) &= 1/12 \end{aligned} \tag{4.20}$$

Compare these results to the results of Jeffrey conditioning:

$$\begin{aligned} P'_{\text{JC}}(E_1) &= 1/2 \\ P'_{\text{JC}}(E_2) &= 3/8 \\ P'_{\text{JC}}(E_3) &= 1/8 \end{aligned} \tag{4.21}$$

Note that (4.16), (4.21), and (4.20) correspond to A, B, C in (4.12).

4.3.3 Triangulating LP and Jeffrey Conditioning

There is an interesting connection between LP conditioning and Jeffrey conditioning as updating methods. Let B be on the zero-sum line between A and C if and only if

$$d(A, C) = d(A, B) + d(B, C) \tag{4.22}$$

where d is the difference measure we are using, so $d(A, B) = \|B - A\|$ for the geometry of reason and $d(A, B) = D_{\text{KL}}(B, A)$ for information geometry. For the geometry of reason (and Euclidean geometry), the zero-sum line between two probability distributions is just what we intuitively think of as a straight line: in Cartesian coordinates, B is on the zero-sum line strictly between A and C if and only if for some $\vartheta \in (0, 1)$, $b_i = \vartheta a_i + (1 - \vartheta)c_i$ and $i = 1, \ldots, n$.

What the zero-sum line looks like for information theory is illustrated in figure 4.5. The reason for the oddity is that the Kullback-Leibler divergence does not obey TRIANGULARITY, an issue that we will address in detail in subsection 4.5.1). Call B a zero-sum point between A and C if (4.22) holds true. For the geometry of

reason, the zero-sum points are simply the points on the straight line between A and C. For information geometry, the zero-sum poins are the boundary points of the set where you can take a shortcut by making a detour, i.e. all points for which $d(A, B) + d(B, C) < d(A, C)$.

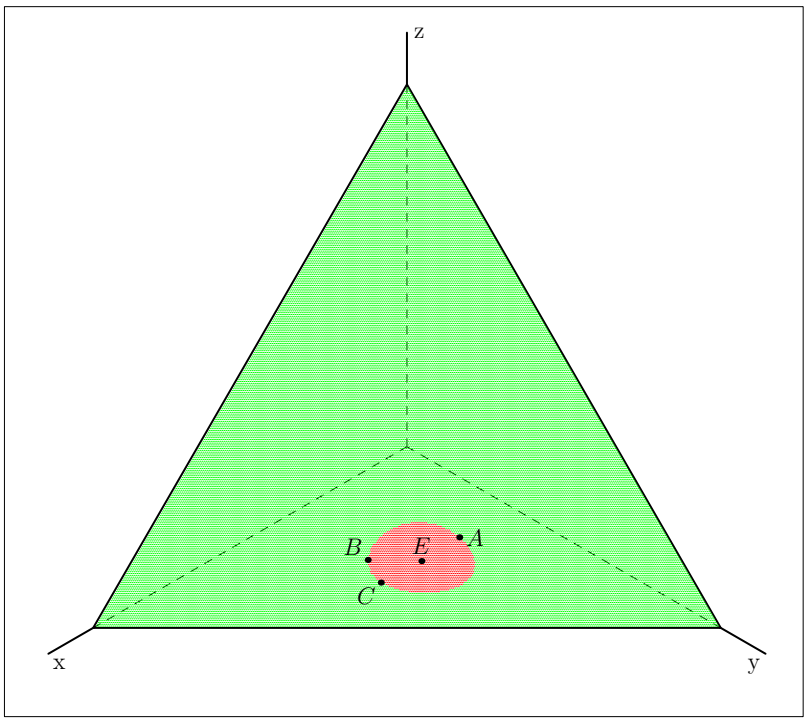

Figure 4.5: The zero-sum line between A and C is the boundary line between the green area, where the triangle inequality holds, and the red area, where the triangle inequality is violated. The posterior probability distribution B recommended by Jeffrey conditioning always lies on the zero-sum line between the prior A and the LP posterior C, as per equation (4.23). E is the point in the red area where the triangle inequality is most efficiently violated.

Remarkably, if A represents a relatively prior probability distribution and C the posterior probability distribution recommended by LP conditioning, the posterior probability distribution recommended by Jeffrey conditioning is always a zero-sum point with respect to the Kullback-Leibler divergence:

$$D_{\text{KL}}(C, A) = D_{\text{KL}}(B, A) + D_{\text{KL}}(C, B) \tag{4.23}$$

Informationally speaking, if you go from A to C, you can just as well go from A to B and then from B to C. This does not mean that we can conceive of information geometry the way we would conceive of non-Euclidean geometry, where it is also possible to travel faster on what from a Euclidean perspective looks like a detour. For in information geometry, you can travel faster on what from the perspective of information theory (!) looks like a detour, i.e. the triangle inequality does not hold.

To prove equation (4.23) in the case $n = 3$ (assuming that LP conditioning does not 'fall off the edge' as in case (b) in Leitgeb and Pettigrew, 2010b, 253) note that all three points (prior, point recommended by Jeffrey conditioning, point recommended by LP conditioning) can be expressed using three variables:

$$\begin{aligned} A &= (1-\alpha, \beta, \alpha-\beta) \\ B &= \left(1-\gamma, \tfrac{\gamma\beta}{\alpha}, \tfrac{\gamma(\alpha-\beta)}{\alpha}\right) \\ C &= \left(1-\gamma, \beta+\tfrac{1}{2}(\gamma-\alpha), \alpha-\beta+\tfrac{1}{2}(\gamma-\alpha)\right) \end{aligned} \tag{4.24}$$

The rest is basic algebra using the definition of the Kullback-Leibler divergence in (3.18). To prove the claim for arbitrary n one simply generalizes (4.24). It is a handy corollary of (4.23) that whenever (A, B) and (B, C) violate TRANSITIVITY OF ASYMMETRY then

$$D_{\text{KL}}(A, C) > D_{\text{KL}}(B, C) + D_{\text{KL}}(A, B) \tag{4.25}$$

in violation of TRIANGULARITY. This way we will not have to go hunting for an example to demonstrate the violation of TRIANGULARITY. A, B, C of (4.12) fulfill

all the conditions for (4.25) and therefore violate TRIANGULARITY.

It is an interesting question to wonder which point E violates the triangle inequality most efficiently so that

$$D_{\mathrm{KL}}(E, C) + D_{\mathrm{KL}}(A, E) \tag{4.26}$$

is minimal. Let $e = (e_1, \ldots, e_n)$ represent E in $\mathbb{S}^{n-1}$. Use the Lagrange Multiplier method to find the Lagrangian

$$\mathcal{L}(e, \lambda) = \sum_{i=1}^{n} e_i \log \frac{e_i}{a_i} + \sum_{i=1}^{n} c_i \log \frac{c_i}{e_i} + \lambda \left(\sum_{i=1}^{n} e_i - 1 \right) \tag{4.27}$$

The Lagrange Multiplier method gives us

$$\frac{\partial \mathcal{L}}{\partial e_k} = \log \frac{e_k}{a_k} + 1 - \frac{r}{q} + \lambda = 0 \text{ for each } k = 1, \ldots, n. \tag{4.28}$$

Manipulate this equation to yield

$$\frac{c_k}{e_k} \exp\left(\frac{c_k}{e_k} \right) = \frac{c_k}{a_k} \exp(1 + \lambda). \tag{4.29}$$

To solve (4.29), use the Lambert W function

$$e_k = \frac{c_k}{W\left(\frac{c_k}{a_k} \exp(1 + \lambda) \right)}. \tag{4.30}$$

Choose λ to fulfill the constraint $\sum e_i = 1$. The result for the discrete case accords with Ovidiu Calin and Constantin Udriste's result for the continuous case (see equation 4.7.9 in Calin and Udriste, 2014, 127). Numerically, for A and C as defined in

equation (4.12),

$$E = (0.415, 0.462, 0.123). \tag{4.31}$$

This is subtly different from the midpoint $m_i = 0.5a_i + 0.5c_i$ (if we were minimizing $D_{\text{KL}}(A, E) + D_{\text{KL}}(C, E)$, the solution would be the midpoint). I do not know whether A, E, C are collinear (see figure 4.5 for illustration).

4.4 Expectations for the Geometry of Reason

This section provides more detail for the expectations in List One (see page 74) and shows how LP conditioning violates them.

4.4.1 Continuity

LP conditioning violates CONTINUITY because standard conditioning gives a different recommendation than a parallel sequence of Jeffrey-type updating scenarios which get arbitrarily close to standard event observation. This is especially troubling considering how important the case for standard conditioning is to Leitgeb and Pettigrew.

To illustrate a CONTINUITY violation, consider the case where Sherlock Holmes reduces his credence that the culprit was male to $\varepsilon_n = 1/n$ for $n = 4, 5, \ldots$. The sequence ε_n is not meant to reflect a case where Sherlock Holmes becomes successively more certain that the culprit was female. It is meant to reflect countably many parallel scenarios which only differ by the degree to which Sherlock Holmes is sure that the culprit was female. These parallel scenarios give rise to a parallel sequence (as opposed to a successive sequence) of updated probabilities $P'_{\text{LP}}(F^{**})$ and another sequence of updated probabilities $P'_{\text{JC}}(F^{**})$ (F^{**} is the proposition that the culprit is female). As $n \to \infty$, both of these sequences go to one.

Straightforward conditionalization on the evidence that 'the culprit is female'

gives us

$$\begin{aligned} P'_{\text{SC}}(E_1) &= 0 \\ P'_{\text{SC}}(E_2) &= 3/4 \\ P'_{\text{SC}}(E_3) &= 1/4. \end{aligned} \tag{4.32}$$

Letting $n \to \infty$ for Jeffrey conditioning yields

$$\begin{aligned} P'_{\text{JC}}(E_1) &= 1/n \quad &\to \quad 0 \\ P'_{\text{JC}}(E_2) &= 3(n-1)/4n \quad &\to \quad 3/4 \\ P'_{\text{JC}}(E_3) &= (n-1)/4n \quad &\to \quad 1/4, \end{aligned} \tag{4.33}$$

whereas letting $n \to \infty$ for LP conditioning yields

$$\begin{aligned} P'_{\text{LP}}(E_1) &= 1/n \quad &\to \quad 0 \\ P'_{\text{LP}}(E_2) &= (4n-3)/6n \quad &\to \quad 2/3 \\ P'_{\text{LP}}(E_3) &= (2n-5)/6n \quad &\to \quad 1/3. \end{aligned} \tag{4.34}$$

LP conditioning violates CONTINUITY.

4.4.2 Regularity

LP conditioning violates REGULARITY because formerly positive probabilities can be reduced to 0 even though the new information in the Jeffrey-type updating scenario makes no such requirements (as is usually the case for standard conditioning). Ironically, Jeffrey-type updating scenarios are meant to be a better reflection of real-life updating because they avoid extreme probabilities.

The violation becomes serious if we are already sympathetic to an information-based account: the amount of information required to turn a non-extreme probability into one that is extreme (0 or 1) is infinite. Whereas the geometry of reason considers extreme probabilities to be easily accessible by non-extreme probabilities under new

information (much like a marble rolling off a table or a bowling ball heading for the gutter), information theory envisions extreme probabilities more like an event horizon. The nearer you are to the extreme probabilities, the more information you need to move on. For an observer, the horizon is never reached.

> **Example 13: Regularity Holmes.** Everything is as in example 11, except that Sherlock Holmes becomes confident to a degree of 2/3 that Mr. R is the culprit and updates his relatively prior probability distribution in (4.16).

Then his posterior probabilities look as follows:

$$\begin{aligned} P'_{\text{JC}}(E_1) &= 2/3 \\ P'_{\text{JC}}(E_2) &= 1/4 \\ P'_{\text{JC}}(E_3) &= 1/12 \end{aligned} \tag{4.35}$$

$$\begin{aligned} P'_{\text{LP}}(E_1) &= 2/3 \\ P'_{\text{LP}}(E_2) &= 1/3 \\ P'_{\text{LP}}(E_3) &= 0 \end{aligned} \tag{4.36}$$

With LP conditioning, Sherlock Holmes' subjective probability that Ms. T is the culprit in example 13 has been reduced to zero. No finite amount of information could bring Ms. T back into consideration as a culprit in this crime, and Sherlock Holmes should be willing to bet any amount of money against a penny that she is not the culprit—even though his evidence is nothing more than an increase in the probability that Mr. R is the culprit.

LP conditioning violates REGULARITY.

4.4.3 Levinstein

LP conditioning violates LEVINSTEIN because of "the potentially dramatic effect [LP conditioning] can have on the likelihood ratios between different propositions" (Levinstein, 2012, 419.). Consider Benjamin Levinstein's example:

Example 14: Levinstein's Ghost. There is a car behind an opaque door, which you are almost sure is blue but which you know might be red. You are almost certain of materialism, but you admit that there's some minute possibility that ghosts exist. Now the opaque door is opened, and the lighting is fairly good. You are quite surprised at your sensory input: your new credence that the car is red is very high.

Jeffrey conditioning leads to no change in opinion about ghosts. Under LP conditioning, however, seeing the car raises the probability that there are ghosts to an astonishing 47%, given Levinstein's reasonable priors. Levinstein proposes a logarithmic inaccuracy measure as a remedy to avoid violation of LEVINSTEIN. As a special case of applying a Levinstein-type logarithmic inaccuracy measure, information theory does not violate LEVINSTEIN.

4.4.4 Invariance

LP conditioning violates INVARIANCE because two agents who have identical credences with respect to a partition of the event space may disagree about this partition after LP conditioning, even when the Jeffrey-type updating scenario provides no new information about the more finely grained partitions on which the two agents disagree.

Example 15: Jane Marple. Jane Marple is on the same case as Sherlock Holmes in example 11 and arrives at the same relatively prior probability distribution as Sherlock Holmes (we will call Jane Marple's relatively prior probability distribution Q and her posterior probability distribution Q'). Jane Marple, however, has a more finely grained probability assignment than Sherlock Holmes and distinguishes between the case where Ms. S went to boarding school with her, of which she has a vague memory, and the case where Ms. S did not and the vague memory is only about a fleeting resemblance of Ms. S with another boarding school mate. Whether or not Ms. S went to boarding school with Jane Marple is completely beside the point with respect to the crime, and Jane Marple considers the possibilities equiprobable whether or not Ms. S went to boarding school with her.

Let $E_2 \equiv E_2^* \vee E_2^{**}$, where E_2^* is the proposition that Ms. S is the culprit and she went to boarding school with Jane Marple and E_2^{**} is the proposition that Ms. S is the culprit and she did not go to boarding school with Jane Marple. Then

$$\begin{aligned} Q(E_1) &= 1/3 \\ Q(E_2^*) &= 1/4 \\ Q(E_2^{**}) &= 1/4 \\ Q(E_3) &= 1/6. \end{aligned} \tag{4.37}$$

Now note that while Sherlock Holmes and Jane Marple agree on the relevant facts of the criminal case (who is the culprit?) in their posterior probabilities if they use Jeffrey conditioning,

$$\begin{aligned} P'_{\text{JC}}(E_1) &= 1/2 \\ P'_{\text{JC}}(E_2) &= 3/8 \\ P'_{\text{JC}}(E_3) &= 1/8 \end{aligned} \tag{4.38}$$

$$\begin{aligned} Q'_{\text{JC}}(E_1) &= 1/2 \\ Q'_{\text{JC}}(E_2^*) &= 3/16 \\ Q'_{\text{JC}}(E_2^{**}) &= 3/16 \\ Q'_{\text{JC}}(E_3) &= 1/8 \end{aligned} \tag{4.39}$$

they do not agree if they use LP conditioning,

$$\begin{aligned} P'_{\text{LP}}(E_1) &= 1/2 \\ P'_{\text{LP}}(E_2) &= 5/12 \\ P'_{\text{LP}}(E_3) &= 1/12 \end{aligned} \tag{4.40}$$

$$\begin{aligned} Q'_{\text{LP}}(E_1) &= 1/2 \\ Q'_{\text{LP}}(E_2^*) &= 7/36 \\ Q'_{\text{LP}}(E_2^{**}) &= 7/36 \\ Q'_{\text{LP}}(E_3) &= 1/9. \end{aligned} \tag{4.41}$$

LP conditioning violates INVARIANCE.

4.4.5 Expansibility

One particular problem with the lack of invariance for LP conditioning is how zero-probability events should be included in the list of prior probabilities that determines the value of the posterior probabilities. Consider

$$\begin{aligned} P(X_1) &= 0 \\ P(X_2) &= 0.3 \\ P(X_3) &= 0.6 \\ P(X_4) &= 0.1 \end{aligned} \tag{4.42}$$

That $P(X_1) = 0$ may be a consequence of standard conditioning in a previous step. Now the agent learns that $P'(X_3 \vee X_4) = 0.5$. Should the agent update on the list presented in (4.42) or on the following list:

$$\begin{aligned} P(X_2) &= 0.3 \\ P(X_3) &= 0.6 \\ P(X_4) &= 0.1 \end{aligned} \tag{4.43}$$

Whether you update on (4.42) or (4.43) makes no difference to Jeffrey conditioning, but due to the lack of invariance it makes a difference to LP conditioning, so the geometry of reason needs to find a principled way to specify the appropriate prior

probabilities. The only non-arbitrary way to do this is either to include or to exclude all zero probability events on the list. This strategy, however, sounds ill-advised unless one signs on to a stronger version of REGULARITY and requires that only a fixed set of events can have zero probabilities (such as logical contradictions), but then the geometry of reason ends up in the catch-22 of LP conditioning running afoul of REGULARITY.

LP conditioning violates EXPANSIBILITY.

4.4.6 Horizon

> **Example 16: Undergraduate Complaint.** An undergraduate student complains to the department head that the professor will not reconsider an 89% grade (which misses an A+ by one percent) when reconsideration was given to other students with a 67% grade (which misses a B- by one percent).

Intuitions may diverge, but the professor's reasoning is as follows. To improve a 60% paper by ten percent is easily accomplished: having your roommate check your grammar, your spelling, and your line of argument will sometimes do the trick. It is incomparably more difficult to improve an 85% paper by ten percent: it may take doing a PhD to turn a student who writes the former into a student who writes the latter. A maiore ad minus, the step from 89% to 90% is greater than the step from 67% to 68%.

Another example for the horizon effect is George Schlesinger's comparison between the risk of a commercial airplane crash and the risk of a military glider landing in enemy territory.

> **Example 17: Airplane Gliders.** Compare two scenarios. In the first, an airplane which is considered safe (probability of crashing is $1/10^9$) goes through an inspection where a mechanical problem is found which increases the probability of a crash to 1/100. In the second, military gliders land behind enemy lines, where their risk of perishing is 26%. A slight change in weather pattern increases this risk to 27%. (Schlesinger, 1995, 211.)

For an interesting instance of the horizon effect in asymmetric multi-dimensional scaling see Chino and Shiraiwa, 1993, section 3, where Naohito Chino and Kenichi Shiraiwa describe as one of the properties of their Hilbert space models of asymmetry how "the similarity between the pair of objects located far from the centroid of objects, say, the origin, is greater than that located near the origin, even if their distances are the same" (42).

I claim that an amujus ought to fulfill the requirements of the horizon effect: it ought to be more difficult to update as probabilities become more extreme (or less middling). I have formalized this requirement in List Two (see page 75). It is trivial that the geometry of reason does not fulfill it. Information theory fails as well, which gives the horizon effect its prominent place in both lists. The way information theory fails, however, is quite different. Near the boundary of $\mathbb{S}^{n-1}$, information theory reflects the horizon effect just as our expectation requires. The problem is near the centre, where some equidistant points are more divergent the closer they are to the middle. I will give an example and more explanation in subsection 4.5.2.

In the next section, I will closely tie the issue of the horizon effect to confirmation. The two main candidates for quantitative measures of relevance confirmation disagree on precisely this issue. Whether you, the reader, will accept the horizon requirement may depend on what your view is on degree of confirmation theory.

4.4.7 Confirmation

The geometry of reason thinks about the comparison of probability distributions in terms of distance. Information theory thinks about the comparison along the lines of information loss when one distribution is used to encode a message rather than the other distribution. One way to test these approaches is to ask how well they align with a third approach to such a comparison: degree of confirmation. Our main concern is the horizon effect of the previous subsection. Which approaches to degree of confirmation theory reflect it, and how do these approaches correspond to the disagreements between information theory and the geometry of reason?

There is, of course, a relevant difference between the aims of the epistemic utility

approach to updating and the aims of degree of confirmation theory. The former investigates norms to which a rational agent conforms in her pursuit of epistemic virtue. The latter seeks to establish qualitative and quantitative measures of impact that evidence has on a hypothesis. Both, however, (I will restrict my attention here to quantitative degree of confirmation theory) attend to the probability of an event, which degree of confirmation theory customarily calls h for hypothesis, before and after the rational agent processes another event, customarily called e for evidence, i.e. $x = P(h|k)$ and $y = P(h|e, k)$ (k is background information).

For perspectives on the link between confirmation and information see Shogenji, 2012, 37f; Crupi and Tentori, 2014; and Milne, 2014, section 4. Vincenzo Crupi and Katya Tentori suggest that there is a "parallelism between confirmation and information search [which] can serve as a valuable heuristic for theoretical work" (Crupi and Tentori, 2014, 89.).

In degree of confirmation theory, incremental confirmation is distinguished from absolute confirmation in the following sense. Let h be the presence of a very rare disease and e a test result such that $y >> x$ but $y < 1-y$. Then, absolutely speaking, e disconfirms h (for Rudolf Carnap, absolute confirmation involves sending y above a threshold r which must be greater than or equal to 0.5). Absolute confirmation is not the subject of this section. I will exclusively discuss incremental confirmation (also called relevance confirmation, just as absolute confirmation is sometimes called firmness confirmation) where $y > x$ implies (incremental) confirmation, $y < x$ implies (incremental) disconfirmation, and $y = x$ implies the lack of both. The difference is illustrated in figure 4.6.

All proposed measures of quantitative, incremental degree of confirmation considered here are a function of x and y. Dependence of incremental confirmation on only x and y is not trivial, as $P(e|k)$ and $P(e|h, k)$ cannot be expressed using only x and y (for a case why dependence should be on only x and y see Atkinson, 2012, 50, with an irrelevant conjunction argument; and Milne, 2014, 254, with a continuity argument). David Christensen's measure $P(h|e, k) - P(h|\neg e, k)$ (see Christensen, 1999, 449) and Robert Nozick's $P(e|h, k) - P(e|\neg h, k)$ (see Nozick, 1981, 252) are not only dependent on x and y, but also on $P(e|k)$, which makes them vulnerable to

Atkinson's and Milne's worries just cited.

Consider the following six contenders for a quantitative, incremental degree of confirmation function, dependent on only x and y. They are based on, in a brief slogan, (i) difference of conditional probabilities, (ii) ratio of conditional probabilities, (iii) difference of odds, (iv) ratio of likelihoods, (v) Gaifman's treatment of Hempel's raven paradox, and (vi) conservation of contrapositivity and commutativity. Logarithms throughout this dissertation are assumed to be the natural logarithm in order to facilitate easy differentiation, although generally a particular choice of base (greater than one) does not make a relevant difference.

$$\begin{aligned}
&\text{(i)} && M_P(x,y) &&= && y - x \\
&\text{(ii)} && R_P(x,y) &&= && \log \frac{y}{x} \\
&\text{(iii)} && J_P(x,y) &&= && \frac{y}{1-y} - \frac{x}{1-x} \\
&\text{(iv)} && L_P(x,y) &&= && \log \frac{y(1-x)}{x(1-y)} \\
&\text{(v)} && G_P(x,y) &&= && \log \frac{1-x}{1-y} \\
&\text{(vi)} && Z_P(x,y) &&= && \begin{cases} \dfrac{y-x}{1-x} & \text{if } y \geq x \\ \dfrac{y-x}{x} & \text{if } y < x \end{cases}
\end{aligned} \tag{4.44}$$

M_P is defended by Carnap, 1962; Earman, 1992; Rosenkrantz, 1994. R_P is defended by Keynes, 1921; Milne, 1996; Shogenji, 2012. J_P is defended by Festa, 1999. L_P is defended by Good, 1950; Good, 1983, chapter 14; Fitelson, 2006; Zalabardo, 2009. G_P is defended by Gaifman, 1979, 120, without the logarithm (I added it to make G_P more comparable to the other functions). Z_P is defended by Crupi et al., 2007.

For more literature supporting the various measures consult footnote 1 in Fitelson, 2001, S124; and an older survey of options in Kyburg, 1983.

To compare how these degree of confirmation measures align with the concept of difference between probability distributions for the purpose of updating it is best to look at derivatives as they reflect the rate of change from the middle to the extremes. This is how we capture the horizon effect requirement for two dimensions. One important difference between degree of confirmation theory and updating is that the former is concerned with a hypothesis and its negation whereas the latter considers all sorts of domains for the probability distribution (in this chapter, I have restricted myself to a finite outcome space). As far as the analogy between degree of confirmation theory on the one hand and updating on the other hand is concerned, we only need to look at the two-dimensional case.

To discriminate between candidates (i)–(vi), I am setting up three criteria (complementing many others in the literature). Let $D(x, y)$ be the generic expression for the degree of confirmation function. Call this List Three.

- ADDITIVITY A theory can be confirmed piecemeal. Whether the evidence is split up into two or more components or left in one piece is irrelevant to the amount of confirmation it confers. Formally, $D(x, z) = D(x, y) + D(y, z)$. Note that this is not the usual triangle inequality because we are in two dimensions.

- SKEW-ANTISYMMETRY It does not matter whether h or $\neg h$ is in view. Confirmation and disconfirmation are commensurable. Formally, $D(x, y) = -D(1 - x, 1 - y)$. A surprising number of candidates fail this requirement, and the requirement is not common in the literature (see, however, the second clause in Milne's fourth desideratum in 1996, 21). In defence of this requirement consider example 18 below. $d_1 > d_2$ may have a negative impact on the latter scientist's grant application, even though the inequality may solely be due to a failure to fulfill skew-antisymmetry.

- CONFIRMATION HORIZON An account of degree of confirmation must exhibit the horizon effect as in List One and List Two, except more simply in two

dimensions. Formally, the functions $\partial D^+_\varepsilon/\partial x$ must be strictly positive and the functions $\partial D^-_\varepsilon/\partial x$ must be strictly negative for all $\varepsilon \in (-1/2, 1/2)$. These functions are defined in (4.45) and (4.46), and I prove in appendix C that the requirement to keep them strictly positive/negative is equivalent to the horizon effect as described formally in List Two (see page 75).

Example 18: Grant Adjudication I. Two scientists compete for grant money. Professor X presents an experiment conferring degree of confirmation d_1 on a hypothesis, if successful; Professor Y presents an experiment conferring degree of disconfirmation $-d_2$ on the negation of the same hypothesis, if unsuccessful. (For the relevance of quantitative confirmation measures to the evaluation of scientific projects see Salmon, 1975, 11.)

The functions for the horizon effect are defined as follows. Let $\varepsilon \in (-1/2, 1/2)$ be fixed. Recall that $D(x, y)$ is the generic expression for a confirmation function measuring the degree of confirmation that a posterior $y = P(h|e, k)$ bestows on a hypothesis for which the prior is $x = P(h|k)$. ε is the difference $y - x$. For $\varepsilon > 0$,

$$\begin{array}{lll} D^-_\varepsilon : (0, \frac{1}{2} - \varepsilon) \to \mathbb{R} & D^-_\varepsilon(x) & = \quad |D(x, x + \varepsilon)| \\ D^+_\varepsilon : (\frac{1}{2}, 1 - \varepsilon) \to \mathbb{R} & D^+_\varepsilon(x) & = \quad |D(x, x + \varepsilon)| \end{array} \tag{4.45}$$

For $\varepsilon < 0$,

$$\begin{array}{lll} D^-_\varepsilon : (-\varepsilon, \frac{1}{2}) \to \mathbb{R} & D^-_\varepsilon(x) & = \quad |D(x, x + \varepsilon)| \\ D^+_\varepsilon : (\frac{1}{2} - \varepsilon, 1) \to \mathbb{R} & D^+_\varepsilon(x) & = \quad |D(x, x + \varepsilon)| \end{array} \tag{4.46}$$

The rate of change for the different quantitative measures of degree of confirmation can be observed in figure 4.6. The pass and fail verdicts in the table below are evident from figure 4.6 and the table of derivatives provided in appendix C. Only J_P, L_P and Z_P fulfill the horizon requirement.

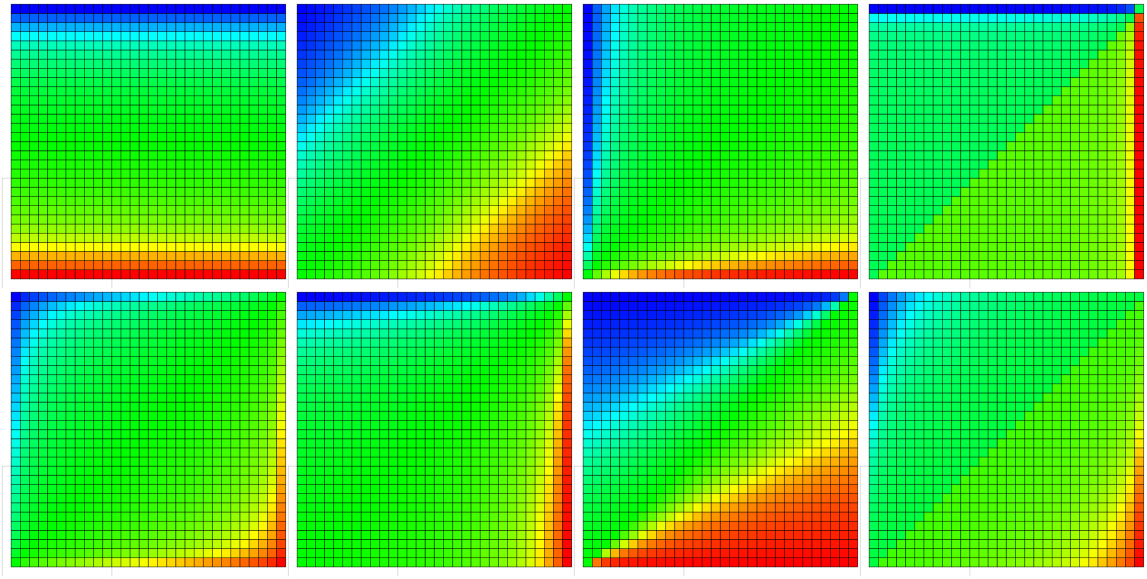

Figure 4.6: Illustration for the six degree of confirmation candidates plus Carnap's firmness confirmation and the Kullback-Leibler divergence. The top row, from left to right, illustrates FMRJ, the bottom row LGZI. 'F' stands for Carnap's firmness confirmation measure $F_P(x, y) = \log(y/(1-y))$. 'M' stands for candidate (i), $M_P(x, y)$ in (4.44), the other letters correspond to the other candidates (ii)-(v). 'I' stands for the Kullback-Leibler divergence multiplied by the sign function of $y - x$ to mimic a quantitative measure of confirmation. For all the squares, the colour reflects the degree of confirmation with x on the x-axis and y on the y-axis, all between 0 and 1. The origin of the square's coordinate system is in the bottom left corner. Blue signifies strong confirmation, red signifies strong disconfirmation, and green signifies the scale between them. Perfect green is $x = y$. G_P looks like it might pass the horizon requirement, but the derivative reveals that it fails CONFIRMATION HORIZON (see appendix C).

Candidate	*Triangularity*	*Skew-Antisymmetry*	*Confirmation Horizon*
M_P	pass	pass	fail
R_P	pass	fail	fail
J_P	pass	fail	pass
L_P	pass	pass	pass
G_P	pass	fail	fail
Z_P	fail	pass	pass

The table makes clear that only L_P passes all three tests. I am not making a strong independent case for L_P here, especially against M_P, which is the most

likely hero of the geometry of reason. This has been done elsewhere (for example in Schlesinger, 1995, where L_P and M_P are compared to each other in their performance given some intuitive examples; Elliott Sober presents the counterargument in Sober, 1994). The argumentative force of this subsection appeals to those who are already sympathetic to L_P. Adherents of M_P will hopefully find other items on List One (see page 74) persuasive and reject the geometry of reason, in which case they may come back to this subsection and re-evaluate their commitment to M_P.

> **Example 19: Grant Adjudication II.** Two scientists compete for grant money. Professor X presents an experiment that will increase the probability of a hypothesis from 98% to 99%, if successful. Professor Y presents an experiment that will increase the probability of a hypothesis from 1% to 2%, if successful.

All else being equal, Professor Y should receive the grant money. If her experiment is more successful, it will arguably make more of a difference. This example illustrates that the analogy between degree of confirmation and updating remains tenuous, since for degree of confirmation theory the consensus on intuitions is far inferior to the updating case. If, for example, the confirmation function is anti-symmetric and $D(x, y) - D(y, x)$ is zero (for M_P and L_P, for example), then together with skew-antisymmetry this means that degree of confirmation is equal for Professor X and Professor Y. Despite its three passes in the above table, L_P fails here.

Based on Roberto Festa's J_P, Professor X's prospective degree of confirmation is 5000 times larger than Professor Y's, but Festa in particular insists "that there is no universally applicable P-incremental c-measure, and that the appropriateness of a P-incremental c-measure is highly context-dependent" (Festa, 1999, 67.). R_P and Z_P appear to be sensitive to example 19. The Kullback-Leibler divergence gives us the right result as well, where the degree of confirmation for going from 1% to 2% is $3.91 \cdot 10^{-3}$ compared to $3.12 \cdot 10^{-3}$ for going from 98% to 99%, but the Kullback-Leibler divergence is not a serious degree of confirmation candidate. It fulfills SKEW-ANTISYMMETRY and CONFIRMATION HORIZON, but not ADDITIVITY (see subsection 4.5.1).

Intuitions easily diverge here. Christensen may be correct when he says, "perhaps the controversy between difference and ratio-based positive relevance models of quantitative confirmation reflects a natural indeterminateness in the basic notion of 'how much' one thing supports another" (Christensen, 1999, 460.). Pluralists allow therefore for "distinct, complementary notions of evidential support" (Hájek and Joyce, 2008, 123.). I am sympathetic towards this indeterminateness in degree of confirmation theory, but not when it comes to updating (see the full employment theorem in section 7.1).

This subsection assumes that despite these problems with the strength of the analogy, degree of confirmation and updating are sufficiently similar to be helpful in associating options with each other and letting the arguments in each other's favour and disfavour cross-pollinate. As an aside, Christensen's S-support given by evidence E is stable over Jeffrey conditioning on $[E, \neg E]$; LP-conditioning is not (see Christensen, 1999, 451). This may serve as another argument from degree of confirmation theory in favour of information theory (which supports Jeffrey conditioning) against the geometry of reason (which supports LP conditioning).

4.5 Expectations for Information Theory

Asymmetry is the central feature of the difference concept that information theory proposes for the purpose of updating between finite probability distributions. In information theory, the information loss differs depending on whether one uses probability distribution P to encode a message distributed according to probability distribution Q, or whether one uses probability distribution Q to encode a message distributed according to probability distribution P. This asymmetry may very well carry over into the epistemic realm. Updating from one probability distribution, for example, which has $P(X) = x > 0$ to $P'(X) = 0$ is common. It is called standard conditioning. Going in the opposite direction, however, from $P(X) = 0$ to $P'(X) = x' > 0$ is controversial and unusual.

The Kullback-Leibler divergence, which is the most promising concept of difference for probability distributions in information theory and the one which gives us

Bayesian standard conditioning as well as Jeffrey conditioning, is non-commutative and may provide the kind of asymmetry required to reflect epistemic asymmetry. However, it also violates TRIANGULARITY, COLLINEAR HORIZON, and TRANSITIVITY OF ASYMMETRY. The task of this section is to show how serious these violations are.

4.5.1 Triangularity

As mentioned at the end of subsection 4.3.3, the three points A, B, C in (4.12) violate TRIANGULARITY as in (4.25):

$$D_{\text{KL}}(A, C) > D_{\text{KL}}(B, C) + D_{\text{KL}}(A, B). \tag{4.47}$$

This is counterintuitive on a number of levels, some of which I have already hinted at in illustration: taking a shortcut while making a detour; buying a pair of shoes for more money than buying the shoes individually.

Information theory, however, does not only violate TRIANGULARITY. It violates it in a particularly egregious way. Consider any distinct two points x and z on $\mathbb{S}^{n-1}$ with coordinates x_i and z_i $(1 \leq i \leq n)$. For simplicity, let us write $\delta(x, z) = D_{\text{KL}}(z, x)$. Then, for any $\vartheta \in (0, 1)$ and an intermediate point y with coordinates $y_i = \vartheta x_i + (1 - \vartheta) z_i$, the following inequality holds true:

$$\delta(x, z) > \delta(x, y) + \delta(y, z). \tag{4.48}$$

I will prove this in a moment, but here is a disturbing consequence: think about an ever more finely grained sequence of partitions y^j, $j \in \mathbb{N}$, of the line segment from x to z with y^{jk} as dividing points. I will spare myself defining these partitions, but note that any dividing point $y^{j_0 k}$ will also be a dividing point in the more finely grained partitions y^{jk} with $j \geq j_0$. Then define the sequence

$$T_j = \sum_k \delta\left(y^{jk}, y^{j(k+1)}\right) \tag{4.49}$$

such that the sum has as many summands as there are dividing points for j, plus one (for example, two dividing points divide the line segment into three possibly unequal thirds). If δ were the Euclidean distance norm, T_j would be constant and would equal $\|z - x\|$. Zeno's arrow moves happily along from x to z, no matter how many stops it makes on the way. Not so for information theory and the Kullback-Leibler divergence. According to (4.48), any stop along the way reduces the sum of divergences.

T_j is a strictly decreasing sequence (does it go to zero? – I do not know, but if yes, it would add to the poignancy of this violation). The more stops you make along the way, the closer you bring together x and z.

For the proof of (4.48), it is straightforward to see that (4.48) is equivalent to

$$\sum_{i=1}^{n}(z_i - x_i)\log\frac{\vartheta x_i + (1-\vartheta)z_i}{x_i} > 0. \tag{4.50}$$

Now we use the following trick. Expand the right hand side to

$$\sum_{i=1}^{n}\left(z_i + \frac{\vartheta}{1-\vartheta}x_i - \frac{\vartheta}{1-\vartheta}x_i - x_i\right)\log\frac{\frac{1}{1-\vartheta}\left(\vartheta x_i + (1-\vartheta)z_i\right)}{\frac{1}{1-\vartheta}x_i} > 0. \tag{4.51}$$

(4.51) is clearly equivalent to (4.50). It is also equivalent to

$$\sum_{i=1}^{n}\left(z_i + \frac{\vartheta}{1-\vartheta}x_i\right)\log\frac{z_i + \frac{\vartheta}{1-\vartheta}x_i}{\frac{1}{1-\vartheta}x_i} + \sum_{i=1}^{n}\frac{1}{1-\vartheta}x_i\log\frac{\frac{1}{1-\vartheta}x_i}{z_i + \frac{\vartheta}{1-\vartheta}x_i} > 0, \tag{4.52}$$

which is true by Gibbs' inequality.

4.5.2 Collinear Horizon

There are two intuitions at work that need to be balanced: on the one hand, the geometry of reason is characterized by simplicity, and the lack of curvature near extreme probabilities may be a price worth paying; on the other hand, simple examples such as example 17 make a persuasive case for curvature.

Information theory is characterized by a very complicated 'semi-quasimetric' (the attribute 'quasi' is due to its non-commutativity, the attribute 'semi' to its violation of the triangle inequality). One of its initial appeals is that it performs well with respect to the horizon requirement near the boundary of the simplex, which is also the location of Schlesinger's examples. It is not trivial, however, to articulate what the horizon requirement really demands.

COLLINEAR HORIZON in List Two seeks to set up the requirement as weakly as possible, only demanding that points collinear with the centre exhibit the horizon effect. The hope is that continuity will take care of the rest, since we want the horizon effect also for probability distributions that are not collinear with the centre. Be that as it may, the Kullback-Leibler divergence fails COLLINEAR HORIZON. Here is a simple example.

$$p = \left(\frac{1}{5}, \frac{2}{5}, \frac{2}{5}\right) \qquad p' = q = \left(\frac{1}{4}, \frac{3}{8}, \frac{3}{8}\right) \qquad q' = \left(\frac{3}{10}, \frac{7}{20}, \frac{7}{20}\right) \tag{4.53}$$

The conditions of COLLINEAR HORIZON in List Two (see page 75) are fulfilled. If p represents A, p' and q represent B, and q' represents C, then note that $\|B - A\| = \|C - B\|$ and M, A, B, C are collinear. In violation of COLLINEAR HORIZON,

$$D_{\mathrm{KL}}(B, A) = 7.3820 \cdot 10^{-3} > 6.4015 \cdot 10^{-3} = D_{\mathrm{KL}}(C, B). \tag{4.54}$$

This violation of an expectation is not as serious as the violation of TRIANGULARITY or TRANSITIVITY OF ASYMMETRY. Just as there is still a reasonable disagreement about difference measures (which do not exhibit the horizon effect) and

ratio measures (which do) in degree of confirmation theory, most of us will not have strong intuitions about the adequacy of information theory based on its violation of COLLINEAR HORIZON. One way in which I can attenuate the independent appeal of this violation against information theory is by making it parasitic on the asymmetry of information theory.

Figure 4.7 illustrates what I mean. Consider the following two inequalities, where M is represented by the centre m of the simplex with $m_i = 1/n$ and Y is an arbitrary probability distribution with X as the midpoint between M and Y, so $x_i = 0.5(m_i + y_i)$.

$$\text{(i) } D_{\text{KL}}(Y, M) > D_{\text{KL}}(M, Y) \text{ and (ii) } D_{\text{KL}}(X, M) > D_{\text{KL}}(Y, X) \tag{4.55}$$

In terms of coordinates, the inequalities reduce to

$$\text{(i) } H(y) < \frac{1}{n} \sum (\log y_i) - \log \frac{1}{n^2} \text{ and} \tag{4.56}$$

$$\text{(ii) } H(y) > \log \frac{4}{n} - \sum \left[\left(\frac{3}{2} y_i + \frac{1}{2n} \right) \log \left(y_i + \frac{1}{n} \right) \right]. \tag{4.57}$$

(i) is simply the case described in the next subsection for asymmetry and illustrated on the bottom left of figure B.1. (ii) tells us how far from the midpoint we can go with a scenario where $p = m, p' = q$ while violating COLLINEAR HORIZON. Clearly, as illustrated in figure 4.7, there is a relationship between asymmetry and COLLINEAR HORIZON.

The bitter aftertaste that remains with COLLINEAR HORIZON is that it is opaque what motivates information theory not only to put probability distributions farther apart near the periphery, as I would expect, but also near the centre. I lack the epistemic intuition reflected in the behaviour. The next subsection on asymmetry deals with this lack of epistemic intuition writ large.

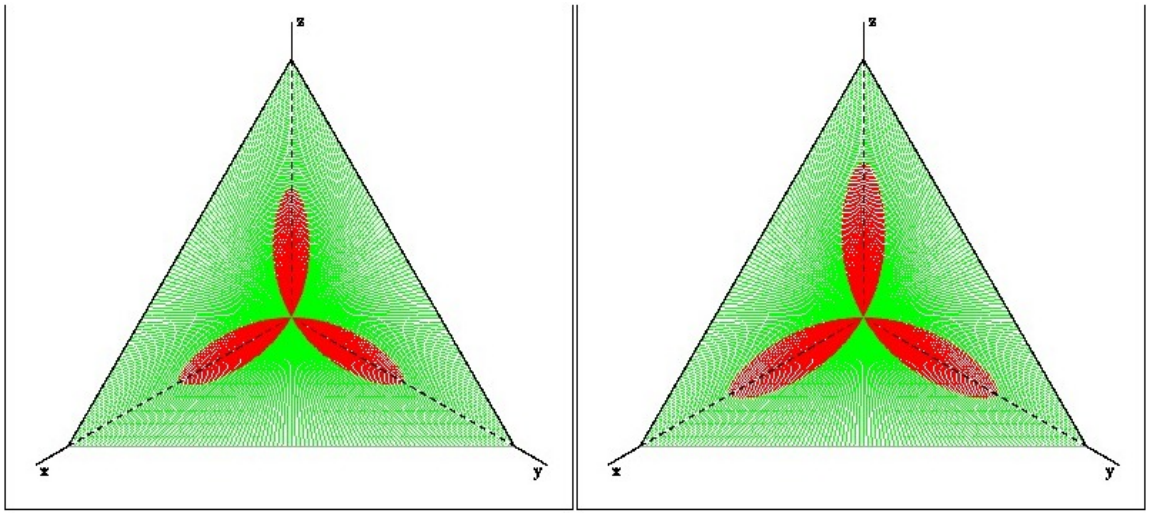

Figure 4.7: These two diagrams illustrate inequalities (4.56) and (4.57). The former displays all points in red which violate COLLINEAR HORIZON, measured from the centre. The latter displays points in different colours whose orientation of asymmetry differs, measured from the centre. The two red sets are not the same, but there appears to be a relationship, one that ultimately I suspect to be due to the more basic property of asymmetry.

4.5.3 Transitivity of Asymmetry

Recall Joyce's two axioms Weak Convexity and Symmetry (see page 73). The geometry of reason (certainly in its Euclidean form) mandates Weak Convexity because the bisector of an isosceles triangle is always shorter than the isosceles sides. Weak Convexity, on the one hand, also holds for information theory (see appendix A for a proof). Symmetry, on the other hand, fails for information theory. Fortunately, although I do not pursue this any further here, information theory arrives at many of Joyce's results even without the violated axiom.

Asymmetry presents a problem for the geometry of reason as well as for information theory. For the geometry of reason, the problem is akin to CONTINUITY. For information theory, the problem is the non-trivial nature of the asymmetries it induces, which somehow need to be reconnected to epistemic justification. I will consider this problem in a moment, but first I will have a look at the problem for the geometry of reason.

Extreme probabilities are special and create asymmetries in updating: moving in

direction from certainty to uncertainty is asymmetrical to moving in direction from uncertainty to certainty. Geometry of reason's metric topology, however, allows for no asymmetries.

> **Example 20: Extreme Asymmetry.** Consider two cases where for case 1 the prior probabilities are $Y_1 = (0.4, 0.3, 0.3)$ and the posterior probabilities are $Y_1' = (0, 0.5, 0.5)$; for case 2 the prior probabilities are reversed, so $Y_2 = (0, 0.5, 0.5)$ and the posterior probabilities $Y_2' = (0.4, 0.3, 0.3)$.

Case 1 is a straightforward application of standard conditioning. Case 2 is more complicated: what does it take to raise a prior probability of zero to a positive number? In terms of information theory, the information required is infinite. Case 2 is also not compatible with standard conditioning (at least not with what Alan Hájek calls the ratio analysis of conditional probability, see Hájek, 2003). The geometry of reason may want to solve this problem by signing on to a version of regularity, but then it violates REGULARITY. Happy kids, clean house, sanity: the hapless homemaker must pick two. The third remains elusive. Continuity, a consistent view of regularity, and symmetry: the hapless geometer of reason cannot have it all.

Now turn to the woes of the information theorist. Given the asymmetric similarity measure of probability distributions that information theory requires (the Kullback-Leibler divergence), a prior probability distribution P may be closer to a posterior probability distribution Q than Q is to P if their roles (prior-posterior) are reversed. That is just what we would expect. The problem is that there is another posterior probability distribution R where the situation is just the opposite: prior P is further away from posterior R than prior R is from posterior P. And whether a probability distribution different from P is of the Q-type or of the R-type escapes any epistemic intuition.

For simplicity, let us consider probability distributions and their associated credence functions on an event space with three atoms $\Omega = \{\omega_1, \omega_2, \omega_3\}$. The simplex $\mathbb{S}^2$ represents all of these probability distributions. Every point p in $\mathbb{S}^2$ representing a probability distribution P induces a partition on $\mathbb{S}^2$ into points that are symmetric

to p, positively skew-symmetric to p, and negatively skew-symmetric to p given the topology of information theory.

In other words, if

$$\Delta_P(P') = D_{\mathrm{KL}}(P', P) - D_{\mathrm{KL}}(P, P'), \tag{4.58}$$

then, holding P fixed, $\mathbb{S}^2$ is partitioned into three regions,

$$\Delta^{-1}(\mathbb{R}_{>0}) \qquad \Delta^{-1}(\mathbb{R}_{<0}) \qquad \Delta^{-1}(\{0\}) \tag{4.59}$$

One could have a simple epistemic intuition such as 'it takes less to update from a more uncertain probability distribution to a more certain probability distribution than the reverse direction,' where the degree of certainty in a probability distribution is measured by its entropy. This simple intuition accords with what we said about extreme probabilities and it holds true for the asymmetric distance measure defined by the Kullback-Leibler divergence in the two-dimensional case where Ω has only two elements (see appendix B).

In higher-dimensional cases, however, the tripartite partition (4.59) is non-trivial—some probability distributions are of the Q-type, some are of the R-type, and it is difficult to think of an epistemic distinction between them that does not already presuppose information theory. See figure B.1 for graphical illustration of this point.

On any account of well-behaved and ill-behaved asymmetries, the Kullback-Leibler divergence is ill-behaved. Of the four axioms as listed by Ralph Kopperman for a distance measure d (see Kopperman, 1988, 89), the Kullback-Leibler divergence violates both symmetry and triangularity, making it a 'semi-quasimetric':

(m1) $d(x, x) = 0$

(m2) $d(x, z) \leq d(x, y) + d(y, z)$ (triangularity)

(m3) $d(x, y) = d(y, x)$ (symmetry)

(m4) $d(x, y) = 0$ implies $x = y$ (separation)

The Kullback-Leibler divergence not only violates symmetry and triangularity, but also TRANSITIVITY OF ASYMMETRY. For a description of TRANSITIVITY OF ASYMMETRY see List Two on page 75. For an example of it, consider

$$P_1 = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right) \qquad P_2 = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \qquad P_3 = \left(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right) \tag{4.60}$$

In the terminology of TRANSITIVITY OF ASYMMETRY in List Two, (P_1, P_2) is asymmetrically positive, and so is (P_2, P_3). The reasonable expectation is that (P_1, P_3) is asymmetrically positive by transitivity, but for the example in (4.60) it is asymmetrically negative.

How counterintuitive this is (epistemically and otherwise) is demonstrated by the fact that in MDS (the multi-dimensional scaling of distance relationships) almost all asymmetric distance relationships under consideration are asymmetrically transitive in this sense, for examples see international trade in Chino, 1978; journal citation in Coombs, 1964; car switch in Harshman et al., 1982; telephone calls in Harshman and Lundy, 1984; interaction or input-output flow in migration, economic activity, and social mobility in Coxon, 1982; flight time between two cities in Gentleman et al., 2006, 191; mutual intelligibility between Swedish and Danish in van Ommen et al., 2013, 193; Tobler's wind model in Tobler, 1975; and the cyclist lovingly hand-sketched in Kopperman, 1988, 91.

This 'ill behaviour' of information theory begs for explanation, or at least classification (it would help, for example, to know that all reasonable non-commutative difference measures used for updating are ill-behaved). Kopperman's objective is primarily to rescue continuity, uniform continuity, Cauchy sequences, and limits for topologies induced by difference measures which violate triangularity, symmetry, and/or separation. Kopperman does not touch axiom (m1), while in the psychological literature (see especially Tversky, 1977) self-similarity is an important topic. This is why an initially promising approach to asymmetric modeling in Hilbert spaces by Chino (see Chino, 1978; Chino, 1990; Chino and Shiraiwa, 1993; and Saburi and

Chino, 2008) will not help us to distinguish well-behaved and ill-behaved asymmetries between probability distributions. I am explaining the reasons in appendix D.

The failure of Chino's modeling approach to make useful distinctions among asymmetric distance measures between probability distributions leads us to the more complex theory of information geometry and differentiable manifolds. Both the results of Shun-ichi Amari (see Shun-ichi, 1985; and Amari and Nagaoka, 2000) and Nikolai Chentsov (see Chentsov, 1982) serve to highlight the special properties of the Kullback-Leibler divergence, not without elevating the discussion to a level of mathematical sophistication, however, where it is difficult to retain the appeal to epistemic intuitions. Information geometry considers probability distributions as differentiable manifolds equipped with a Riemannian metric. This metric, however, is Fisher's information metric, not the Kullback-Leibler divergence, and it is defined on the tangent space of the simplex representing finite-dimensional probability distributions. There is a sense in which the Fisher information metric is the derivative of the Kullback-Leibler divergence, and so the connection to epistemic intuitions can be re-established.

For a future research project, it would be lovely either to see information theory debunked in favour of an alternative geometry (this chapter has demonstrated that this alternative will not be the geometry of reason); or to see uniqueness results for the Kullback-Leibler divergence to show that despite its ill behaviour the Kullback-Leibler is the right asymmetric distance measure on which to base inference and updating. Chentsov's theory of monotone invariance and Amari's theory of α-connections are potential candidates to provide such results as well as an epistemic justification for information theory.

Leitgeb and Pettigrew's reasoning to establish LP conditioning on the basis of the geometry of reason is valid. Given the failure of LP conditioning with respect to expectations in List One, it cannot be sound. The premise to reject is the geometry of reason. A competing approach, information theory, yields results that fulfill all of these expectations except HORIZON. Information theory, however, fails two other expectations identified in List Two—expectations which the geometry of reason fulfills. We are left with loose ends and ample opportunity for further work.

The epistemic utility approach, itself a relatively recent phenomenon, needs to come to a deeper understanding of its relationship with information theory. It is an open question, for example, if it is possible to provide a complete axiomatization consistent with information theory to justify probabilism, standard conditioning, and Jeffrey conditioning from an epistemic utility approach as Shore and Johnston have done from a pragmatic utility approach. It is also an open question, given the results of this chapter, if there is hope reconciling information theory with intuitions we have about epistemic utility and its attendant quantitative concept of difference for partial beliefs.

Chapter 5

A Natural Generalization of Jeffrey Conditioning

5.1 Two Generalizations

Carl Wagner presents a case where Jeffrey conditioning does not apply but intuition dictates a solution in keeping with the same principle that motivates Jeffrey conditioning (see Wagner, 1992). I will sometimes call this intuition (W) for short, but usually I will call it Jeffrey's Updating Principle JUP. Wagner's (W) solution, or Wagner conditioning, generalizes Jeffrey conditioning. PME, of course, also generalizes Jeffrey conditioning. Wagner investigates whether his generalization (W) agrees with PME, finding that it does not. He then uses his method not only to present a "natural generalization of Jeffrey conditioning" (see Wagner, 1992, 250), but also to deepen criticism of PME.

I will show that PME not only generalizes Jeffrey conditioning (as is well known, for a formal proof see Caticha and Giffin, 2006) but also Wagner conditioning. Wagner's intuition (W) is plausible, and his method works. His derivation of a disagreement with PME, however, is conceptually more complex than he assumes. Specifically, his derivation relies on an assumption that is rejected by advocates of PME: that rational agents can (indeed, should) have imprecise credences. In this chapter, I show that if agents have sharp credences, Wagner conditioning actually agrees with PME. In chapter 6, I examine the debate about imprecise credences.

Below, we will show that PME and (W) are consistent given (L). (L) is what I call the Laplacean principle which requires a rational agent, besides other standard Bayesian commitments, to hold sharp credences with respect to well-defined events

under consideration. (I), which is inconsistent with (L) and which some Bayesians accept, allows a rational agent to have indeterminate or imprecise credences (see Ellsberg, 1961; Levi, 1985; Walley, 1991; and Joyce, 2010). The following table summarizes the positions in diagram form. '✓' means that the positions with a mark ('•') are consistent with each other; '×' means that they are inconsistent with each other.

PME	(W)	(I)	(L)		
•	•			×	according to Wagner's article
•	•			✓	according to this article
		•	•	×	disagree over permitting indeterminacy
•	•	•		×	formally shown in Wagner's article
•	•		•	✓	formally shown here

While Wagner is welcome to deny (L), my sense is that advocates of PME usually accept it because they are already to the moderate right of Sandy L. Zabell's spectrum between left-wing dadaists and right-wing totalitarians (see Zabell, 2005, 27; Zabell's representative of right-wing totalitarianism is E.T. Jaynes). If there were an advocate of PME sympathetic to (I), Wagner's result would indeed force her to choose. Wagner's criticism of PME is misplaced, however, since it rests on a hidden assumption that someone who believes in PME would tend not to hold. Wagner does not give an independent argument for (I). This chapter shows how elegantly PME generalizes not only standard conditioning and Jeffrey conditioning but also Wagner conditioning, once we accept (L). In chapter 6, I provide independent reasons why we should accept (L).

A tempered and differentiated account of PME (contrasted with Jaynes' right-wing earlier version) is not only largely immune to criticisms, but often illuminates the problems that the criticisms pose (for example the *Judy Benjamin* case; see chapter 7). This account rests on principles such as (L) and a reasonable interpretation of what we mean by objectivity (see my comments about Donkin's objectivism in section 2.1). To make this more clear, and before we launch into the formalities of generalizing Wagner conditioning by using PME, let us articulate (L). (L) is what I call the Laplacean principle and in addition to standard Bayesian com-

mitments states that a rational agent assigns a determinate precise probability to a well-defined event under consideration (for a defence of (L) against (I) see White, 2010; and Elga, 2010). (I) is the negation of (L): rational agents sometimes do not assign determinate precise probabilities to a well-defined event under consideration.

To avoid excessive apriorism (see Seidenfeld, 1979), (L) does not require that a rational agent has probabilities assigned to all events in an event space, only that, once an event has been brought to attention, and sometimes retrospectively, the rational agent is able to assign a sharp probability (for more detail on this issue and some rules about which propositions need to be assigned subjective probabilities see Hájek, 2003, 313). Newton did not need to have a prior probability for Einstein's theory in order to have a posterior probability for his theory of gravity.

(L) also does not require objectivity in the sense that all rational agents must agree in their probability distributions if they have the same information. It is important to distinguish between type I and type II prior probabilities (I have also been calling them absolutely and relatively prior probabilities). The former precede any information at all (so-called ignorance priors). The latter are simply prior relative to posterior probabilities in probability kinematics. They may themselves be posterior probabilities with respect to an earlier instance of probability kinematics.

The case for objectivity in probability kinematics, where prior probabilities are of type II, is consistent with and dependent on a subjectivist interpretation of probabilities, making for some terminological confusion. Interpretations of the evidence and how it is to be cast in terms of formal constraints may vary. Once we agree on a prior distribution (type II), however, and on a set of formal constraints representing our evidence, PME claims that posterior probabilities follow mechanically. Just as is the case in deductive logic, we may come to a tentative and voluntary agreement on an interpretation, a set of rules and presuppositions and then go part of the way together. For the interplay between PME and *Infomin*, see section 3.2.

Some advocates of PME may find (L) too weak in its claims, but none think it is too strong. Once (L) is assumed, however, Wagner's diagnosis of disagreement between (W) and PME fails. Moreover, PME and (L) together seamlessly generalize Wagner conditioning. In the remainder of this chapter I will provide a sketch of

a formal proof for this claim. I make the case for (L) in chapter 6. A welcome side-effect of reinstating harmony between PME and (W) is that it provides an inverse procedure to Vladimír Majerník's method of finding marginals based on given conditional probabilities (see Majerník, 2000; and section 5.4).

5.2 Wagner's Natural Generalization of Jeffrey Conditioning

Wagner claims that he has found a relatively common case of probability kinematics in which PME delivers the wrong result so that we must develop an ad hoc generalization of Jeffrey conditioning. This is best explained by using Wagner's example, the *Linguist* problem (see Wagner, 1992, 252 and Spohn, 2012, 197).

> **Example 21: Linguist.** You encounter the native of a certain foreign country and wonder whether he is a Catholic northerner (θ_1), a Catholic southerner (θ_2), a Protestant northerner (θ_3), or a Protestant southerner (θ_4). Your prior probability p over these possibilities (based, say, on population statistics and the judgment that it is reasonable to regard this individual as a random representative of his country) is given by $p(\theta_1) = 0.2, p(\theta_2) = 0.3, p(\theta_3) = 0.4,$ and $p(\theta_4) = 0.1$. The individual now utters a phrase in his native tongue which, due to the aural similarity of the phrases in question, might be a traditional Catholic piety (ω_1), an epithet uncomplimentary to Protestants (ω_2), an innocuous southern regionalism (ω_3), or a slang expression used throughout the country in question (ω_4). After reflecting on the matter you assign subjective probabilities $u(\omega_1) = 0.4, u(\omega_2) = 0.3, u(\omega_3) = 0.2,$ and $u(\omega_4) = 0.1$ to these alternatives. In the light of this new evidence how should you revise p?

Let $\Theta = \{\theta_i : i = 1, \ldots, 4\}, \Omega = \{\omega_i : i = 1, \ldots, 4\}$. Let $\Gamma : \Omega \to 2^{\Theta} - \{\emptyset\}$ be the function which maps ω to $\Gamma(\omega)$, the narrowest event in Θ entailed by the outcome $\omega \in \Omega$. Here are two definitions that take advantage of the apparatus established by Arthur Dempster (see Dempster, 1967). We will need m and b to articulate Wagner's (W) solution for *Linguist* type problems.

$$\text{For all } E \subseteq \Theta, m(E) = u(\{\omega \in \Omega : \Gamma(\omega) = E\}). \tag{5.1}$$

$$\text{For all } E \subseteq \Theta, b(E) = \sum_{H \subseteq E} m(H) = u(\{\omega \in \Omega : \Gamma(\omega) \subseteq E\}). \tag{5.2}$$

Let Q be the posterior joint probability measure on $\Theta \times \Omega$, and Q_Θ the marginalization of Q to Θ, Q_Ω the marginalization of Q to Ω. Wagner plausibly suggests that Q is compatible with u and Γ if and only if

$$\text{for all } \theta \in \Theta \text{ and for all } \omega \in \Omega, \theta \notin \Gamma(\omega) \text{ implies that } Q(\theta, \omega) = 0 \tag{5.3}$$

and

$$Q_\Omega = u. \tag{5.4}$$

The two conditions (5.3) and (5.4), however, are not sufficient to identify a "uniquely acceptable revision of a prior" (Wagner, 1992, 250). Wagner's proposal includes a third condition, which extends Jeffrey's rule to the situation at hand. We will call it (W). To articulate the condition, we need some more definitions. For all $E \subseteq \Theta$, let $E_\bigstar = \{\omega \in \Omega : \Gamma(\omega) = E\}$, so that $m(E) = u(E_\bigstar)$. For all $A \subseteq \Theta$ and all $B \subseteq \Omega$, let "A" $= A \times \Omega$ and "B" $= \Theta \times B$, so that $Q(\text{"A"}) = Q_\Theta(A)$ for all $A \subseteq \Theta$ and $Q(\text{"B"}) = Q_\Omega(B)$ for all $B \subseteq \Omega$. Let also $\mathcal{E} = \{E \subseteq \Theta : m(E) > 0\}$ be the family of evidentiary focal elements.

According to Wagner only those Q satisfying the condition

$$\text{for all } A \subseteq \Theta \text{ and for all } E \in \mathcal{E}, Q(\text{"A"} \,|\, \text{"E}_\bigstar\text{"}) = p(A|E) \tag{5.5}$$

are eligible candidates for updated joint probabilities in *Linguist* type problems. To adopt (5.5), says Wagner, is to make sure that the total impact of the occurrence of the event $E_\bigstar$ is to preclude the occurrence of any outcome $\theta \notin E$, and that, within E, p remains operative in the assessment of relative uncertainties (see Wagner, 1992, 250). While conditions (5.3), (5.4) and (5.5) may admit an infinite number of joint probability distributions on $\Theta \times \Omega$, their marginalizations to Θ are identical and give us the desired posterior probability, expressible by the formula

$$q(A) = \sum_{E \in \mathcal{E}} m(E)p(A|E). \tag{5.6}$$

So far we are in agreement with Wagner. Wagner's scathing verdict about PME towards the end of his article, however, is not really a verdict about PME in the Laplacean tradition but about the curious conjunction of PME and (I). Wagner never invokes (I); it remains an enthymematic assumption in Wagner's argument.

> Students of maximum entropy approaches to probability revision may [...] wonder if the probability measure defined by our formula (5.6) similarly minimizes [the Kullback-Leibler information number] $D_{\mathrm{KL}}(q, p)$ over all probability measures q bounded below by b. The answer is negative [...] convinced by Skyrms, among others, that PME is not a tenable updating rule, we are undisturbed by this fact. Indeed, we take it as additional evidence against PME that (5.6), firmly grounded on [...] a considered judgment that (5.5) holds, might violate PME [...] the fact that Jeffrey's rule coincides with PME is simply a misleading fluke, put in its proper perspective by the natural generalization of Jeffrey conditioning described in this paper. [References to formulas and notation modified.] (Wagner, 1992, 255.)

In the next section, we will contrast what Wagner considers to be the solution of PME for this problem, 'Wagner's PME solution,' and Wagner's solution presented in this section, 'Wagner's (W) solution,' and show, in much greater detail than Wagner does, why Wagner's PME solution misrepresents PME.

5.3 Wagner's PME Solution

Wagner's PME solution assumes the constraint that b must act as a lower bound for the posterior probability. Consider $E_{12} = \{\theta_1 \vee \theta_2\}$. Because both ω_1 and ω_2 entail E_{12}, according to (5.2), $b(E_{12}) = 0.70$. It makes sense to consider it a constraint that the posterior probability for E_{12} must be at least $b(E_{12})$. Then we choose from all probability distributions fulfilling the constraint the one which is closest to the prior probability distribution, using the Kullback-Leibler divergence.

Wagner applies this idea to the marginal probability distribution on Θ. He does not provide the numbers, but refers to simpler examples to make his point that PME does not generally agree with his solution. To aid the discussion, I want to populate Wagner's claim for the *Linguist* problem with numbers. Using proposition 1.29 in Dimitri Bertsekas' book *Constrained Optimization and Lagrange Multiplier Methods* (see Bertsekas, 1982, 71) and some non-trivial calculations, Wagner's PME solution for the *Linguist* problem (indexed Q_{wm}) is

$$\tilde{\beta} = (Q_{wm}(\theta_j))^\intercal = (0.30, 0.45, 0.10, 0.15)^\intercal. \tag{5.7}$$

A brief remark about notation: I will use α for vectors expressing ω_i probabilities and β for vectors expressing θ_j probabilities. I will use a tilde as in $\tilde{\beta}$ or a hat as in $\hat{\beta}$ for posteriors, while priors remain without such ornamentation. The tilde is used for Wagner's PME solution (which, as we will see, is incorrect) and the hat for the correct solution (both (W) and PME).

The cross-entropy between $\tilde{\beta}$ and the prior

$$\beta = (P(\theta_j))^\intercal = (0.20, 0.30, 0.40, 0.10)^\intercal \tag{5.8}$$

is indeed significantly smaller than the cross-entropy between Wagner's (W) solution

$$\hat{\beta} = (Q(\theta_j))^{\intercal} = (0.30, 0.60, 0.04, 0.06)^{\intercal} \tag{5.9}$$

and the prior β (0.0823 compared to 0.4148). For the cross-entropy, we use the Kullback-Leibler Divergence

$$D_{\text{KL}}(q, p) = \sum_j q(\theta_j) \log_2 \frac{q(\theta_j)}{p(\theta_j)}. \tag{5.10}$$

From the perspective of a PME advocate, there are only two explanations for this difference in cross-entropy. Either Wagner's (W) solution illegitimately uses information not contained in the problem, or Wagner's PME solution has failed to include information that is contained in the problem. I will simplify the *Linguist* problem in order to show that the latter is the case.

> **Example 22: Simplified Linguist.** Imagine the native is either Protestant or Catholic (50:50). Further imagine that the utterance of the native either entails that the native is a Protestant (60%) or provides no information about the religious affiliation of the native (40%).

Using (5.6), the posterior probability distribution is 80:20 (Wagner's (W) solution and, surely, the correct solution). Using b as a lower bound and PME, Wagner's PME solution for this radically simplified problem is 60:40, clearly a more entropic solution than Wagner's (W) solution. The problem, as we will show, is that Wagner's PME solution does not take into account (L), which an PME advocate would naturally accept.

For a Laplacean, the prior joint probability distribution on $\Theta \times \Omega$ is not left unspecified for the calculation of the posteriors. Before the native makes the utterance, the event space is unspecified with respect to Ω. After the utterance, however, the event space is defined (or brought to attention) and populated by prior probabilities according to (L). That this happens retrospectively may or may not be a problem:

Bayes' theorem is frequently used retrospectively, for example when the anomalous precession of Mercury's perihelion, discovered in the mid-1800s, was used to confirm Albert Einstein's General Theory of Relativity in 1915 (for retrospective conditioning see Grove and Halpern, 1997; and Diaconis and Zabell, 1982, 822). I shall bracket for now that this procedure is controversial and refer the reader to the literature on Old Evidence.

Ariel Caticha and Adom Giffin make the following appeal:

> Bayes' theorem requires that $P(\omega, \theta)$ be defined and that assertions such as 'ω *and* θ' be meaningful; the relevant space is neither Ω nor Θ but the product $\Omega \times \Theta$ [notation modified] (Caticha and Giffin, 2006, 9.)

Following (L) we shall populate the joint probability matrix P on $\Omega \times \Theta$, which is a matter of synchronic constraints, as updating the joint probability P to Q on $\Omega \times \Theta$ is a matter of diachronic constraints. For the *Simplified Linguist* problem, this procedure gives us the correct result, agreeing with Wagner's (W) solution (80:20).

There is a more general theorem which incorporates Wagner's (W) method into Laplacean realism and PME orthodoxy. The proof of this theorem is in the next section. Its validity is confirmed by how well it works for the *Linguist* problem (as well as the *Simplified Linguist* problem).

We have not yet formally demonstrated that for all Wagner-type problems $(\beta, \hat{\alpha}, \kappa)$, the correct PME solution (versus Wagner's deficient PME solution) agrees with Wagner's (W) solution, although we have established a useful framework and demonstrated the agreement for the *Linguist* problem. As Vladimír Majerník has shown how to derive marginal probabilities from conditional probabilities using PME (see Majerník, 2000), in the next section I will inversely show how to derive conditional probabilities (i.e. the joint probability matrices) from the marginal probabilities and logical relationships provided in Wagner-type problems. This technical result together with the claim established in the present paper that Wagner's intuition (W) is consistent with PME, given (L), underlines the formal and conceptual virtue of PME.

5.4 Maximum Entropy and Probability Kinematics Constrained by Conditionals

Jeffrey conditioning is a method of update (recommended first by Richard Jeffrey in Jeffrey, 1965) which generalizes standard conditioning and operates in probability kinematics where evidence is uncertain ($P(E) \neq 1$). Sometimes, when we reason inductively, outcomes that are observed have entailment relationships with partitions of the possibility space that pose challenges that Jeffrey conditioning cannot meet. As we will see, it is not difficult to resolve these challenges by generalizing Jeffrey conditioning. There are claims in the literature that the principle of maximum entropy, from now on PME, conflicts with this generalization. I will show under which conditions this conflict obtains. Since proponents of PME are unlikely to subscribe to these conditions, the position of PME in the larger debate over inductive logic and reasoning is not undermined.

In his paper "Marginal Probability Distribution Determined by the Maximum Entropy Method" (see Majerník, 2000), Vladimír Majerník asks the following question: If we had two partitions of an event space and knew all the conditional probabilities (any conditional probability of one event in the first partition conditional on another event in the second partition), would we be able to calculate the marginal probabilities for the two partitions? The answer is yes, if we commit ourselves to PME.

For Majerník's question, PME provides us with a unique and plausible answer. We may also be interested in the obverse question: if the marginal probabilities of the two partitions were given, would we similarly be able to calculate the conditional probabilities? The answer is yes: given PME, Theorems 2.2.1. and 2.6.5. in Cover and Thomas, 2006 reveal that the joint probabilities are the product of the marginal probabilities (see also Debbah and Müller, 2005). Once the joint probabilities and the marginal probabilities are available, it is trivial to calculate the conditional probabilities.

It is important to note that these joint probabilities do not legislate independence, even though they allow it. Mérouane Debbah and Ralf Müller correctly describe these

joint probabilities as a model with as many degrees of freedom as possible, which leaves free degrees for correlation to exist or not (see Debbah and Müller, 2005, 1674).

I will now present Jeffrey conditioning in unusual notation, anticipating using this notation to solve Wagner's *Linguist* problem and to give a general solution for the obverse Majerník problem. Let Ω be a finite event space and $\{\theta_j\}_{j=1,\ldots,n}$ a partition of Ω. Let κ be an $m \times n$ matrix for which each column contains exactly one 1, otherwise 0. Let $P = P_{\text{prior}}$ and $\hat{P} = P_{\text{posterior}}$. Then $\{\omega_i\}_{i=1,\ldots,m}$, for which

$$\omega_i = \bigcup_{j=1,\ldots,n} \theta^*_{ij}, \tag{5.11}$$

is likewise a partition of Ω (the ω are basically a more coarsely grained partition than the θ). $\theta^*_{ij} = \emptyset$ if $\kappa_{ij} = 0$, $\theta^*_{ij} = \theta_j$ otherwise. Let β be the vector of prior probabilities for $\{\theta_j\}_{j=1,\ldots,n}(P(\theta_j) = \beta_j)$ and $\hat{\beta}$ the vector of posterior probabilities ($\hat{P}(\theta_j) = \hat{\beta}_j$); likewise for α and $\hat{\alpha}$ corresponding to the prior and posterior probabilities for $\{\omega_i\}_{i=1,\ldots,m}$, respectively.

A Jeffrey-type problem is when β and $\hat{\alpha}$ are given and we are looking for $\hat{\beta}$. A mathematically more concise characterization of a Jeffrey-type problem is the triple $(\kappa, \beta, \hat{\alpha})$. The solution, using Jeffrey conditioning, is

$$\hat{\beta}_j = \beta_j \sum_{i=1}^{n} \frac{\kappa_{ij}\hat{\alpha}_i}{\sum_{l=1}^{m} \kappa_{il}\beta_l} \text{ for all } j = 1, \ldots, n. \tag{5.12}$$

The notation is more complicated than it needs to be for Jeffrey conditioning. In section 5, however, I will take full advantage of it to present a generalization where the ω_i do not range over the θ_j. In the meantime, here is an example to illustrate (5.12).

> **Example 23: Colour Blind.** A token is pulled from a bag containing 3 yellow tokens, 2 blue tokens, and 1 purple token. You are colour blind and cannot distinguish between the blue and the purple token when you see it. When the token is pulled, it

is shown to you in poor lighting and then obscured again.

You come to the conclusion based on your observation that the probability that the pulled token is yellow is 1/3 and that the probability that the pulled token is blue or purple is 2/3. What is your updated probability that the pulled token is blue? Let $P(\text{blue})$ be the prior subjective probability that the pulled token is blue and $\hat{P}(\text{blue})$ the respective posterior subjective probability. Jeffrey conditioning, based on JUP (which mandates, for example, that $\hat{P}(\text{blue}|\text{blue or purple}) = P(\text{blue}|\text{blue or purple})$) recommends

$$\begin{aligned}\hat{P}(\text{blue}) \quad =&\hat{P}(\text{blue}|\text{blue or purple})\hat{P}(\text{blue or purple})+ \\ &\hat{P}(\text{blue}|\text{neither blue nor purple})\hat{P}(\text{neither blue nor purple}) \\ =&P(\text{blue}|\text{blue or purple})\hat{P}(\text{blue or purple}) = 4/9 \end{aligned} \tag{5.13}$$

In the notation of (5.12), example 23 is calculated with $\beta = (1/2, 1/3, 1/6)^\top, \hat{\alpha} = (1/3, 2/3)^\top$,

$$\kappa = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} \tag{5.14}$$

and yields the same result as (5.13) with $\hat{\beta}_2 = 4/9$.

Wagner's *Linguist* problem is an instance of the more general obverse Majerník problem where partitions are given with logical relationships between them as well as some marginal probabilities. Wagner-type problems seek as a solution missing marginals, while obverse Majerník problems seek the conditional probabilities as well, both of which I will eventually provide using PME.

In order to show how PME generalizes Jeffrey conditioning (see subsection 3.2.5) and Wagner conditioning to boot, I use the notation that I have already introduced for Jeffrey conditioning. We can characterize Wagner-type problems analogously to Jeffrey-type problems by a triple $(\kappa, \beta, \hat{\alpha})$. $\{\theta_j\}_{j=1,\ldots,n}$ and $\{\omega_i\}_{i=1,\ldots,m}$ now refer

to independent partitions of Ω, i.e. (5.11) need not be true. Besides the marginal probabilities $P(\theta_j) = \beta_j, \hat{P}(\theta_j) = \hat{\beta}_j, P(\omega_i) = \alpha_i, \hat{P}(\omega_i) = \hat{\alpha}_i$, we therefore also have joint probabilities $\mu_{ij} = P(\omega_i \cap \theta_j)$ and $\hat{\mu}_{ij} = \hat{P}(\omega_i \cap \theta_j)$.

Given the specific nature of Wagner-type problems, there are a few constraints on the triple $(\kappa, \beta, \hat{\alpha})$. The last row $(\mu_{mj})_{j=1,\ldots,n}$ is special because it represents the probability of ω_m, which is the negation of the events deemed possible after the observation. In the *Linguist* problem, for example, ω_5 is the event (initially highly likely, but impossible after the observation of the native's utterance) that the native does not make any of the four utterances. The native may have, after all, uttered a typical Buddhist phrase, asked where the nearest bathroom was, complimented your fedora, or chosen to be silent. κ will have all 1s in the last row. Let $\hat{\kappa}_{ij} = \kappa_{ij}$ for $i = 1, \ldots, m-1$ and $j = 1, \ldots, n$; and $\hat{\kappa}_{mj} = 0$ for $j = 1, \ldots, n$. $\hat{\kappa}$ equals κ except that its last row are all 0s, and $\hat{\alpha}_m = 0$. Otherwise the 0s are distributed over κ (and equally over $\hat{\kappa}$) so that no row and no column has all 0s, representing the logical relationships between the ω_is and the θ_js ($\kappa_{ij} = 0$ if and only if $\hat{P}(\omega_i \cap \theta_j) = \mu_{ij} = 0$). We set $P(\omega_m) = x$ ($\hat{P}(\omega_m) = 0$), where x depends on the specific prior knowledge. Fortunately, the value of x cancels out nicely and will play no further role. For convenience, we define

$$\zeta = (0, \ldots, 0, 1)^\top \tag{5.15}$$

with $\zeta_m = 1$ and $\zeta_i = 0$ for $i \neq m$.

The best way to visualize such a problem is by providing the joint probability matrix $M = (\mu_{ij})$ together with the marginals α and β in the last column/row, here for example as for the *Linguist* problem with $m = 5$ and $n = 4$ (note that this is not the matrix M, which is $m \times n$, but M expanded with the marginals in improper matrix notation):

$$\begin{bmatrix} \mu_{11} & \mu_{12} & 0 & 0 & \alpha_1 \\ \mu_{21} & \mu_{22} & 0 & 0 & \alpha_2 \\ 0 & \mu_{32} & 0 & \mu_{34} & \alpha_3 \\ \mu_{41} & \mu_{42} & \mu_{43} & \mu_{44} & \alpha_4 \\ \mu_{51} & \mu_{52} & \mu_{53} & \mu_{54} & x \\ \beta_1 & \beta_2 & \beta_3 & \beta_4 & 1.00 \end{bmatrix}. \tag{5.16}$$

The $\mu_{ij} \neq 0$ where $\kappa_{ij} = 1$. Ditto, mutatis mutandis, for $\hat{M}, \hat{\alpha}, \hat{\beta}$. To make this a little less abstract, Wagner's *Linguist* problem is characterized by the triple $(\kappa, \beta, \hat{\alpha})$,

$$\kappa = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \text{ and } \hat{\kappa} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{5.17}$$

$$\beta = (0.2, 0.3, 0.4, 0.1)^\top \text{ and } \hat{\alpha} = (0.4, 0.3, 0.2, 0.1, 0)^\top. \tag{5.18}$$

Wagner's solution, based on JUP, is

$$\hat{\beta}_j = \beta_j \sum_{i=1}^{m-1} \frac{\hat{\kappa}_{ij}\hat{\alpha}_i}{\sum_{\hat{\kappa}_{il}=1} \beta_l} \text{ for all } j = 1, \ldots, n. \tag{5.19}$$

In numbers,

$$\hat{\beta}_j = (0.3, 0.6, 0.04, 0.06)^\top. \tag{5.20}$$

The posterior probability that the native encountered by the linguist is a northerner, for example, is 34%. Wagner's notation is completely different and never specifies or provides the joint probabilities, but I hope the reader appreciates both the analogy to (5.12) underlined by this notation as well as its efficiency in delivering a correct PME solution for us. The solution that Wagner attributes to PME is misleading because of Wagner's Dempsterian setup which does not take into account that proponents of PME are likely to be proponents of the classical Bayesian position that type II prior probabilities are specified and determinate once the agent attends to the events in question. Some Bayesians in the current discussion explicitly disavow this requirement for (possibly retrospective) determinacy (especially James Joyce in 2010 and other papers). Proponents of PME (a proper subset of Bayesians), however, are unlikely to follow Joyce—if they did, they would indeed have to address Wagner's example to show that their allegiances to PME and to indeterminacy are compatible.

That (5.19) follows from JUP is well-documented in Wagner's paper. For PME solution for this problem, I will not use (5.19) or JUP, but maximize the entropy for the joint probability matrix M and then minimize the cross-entropy between the prior probability matrix M and the posterior probability matrix $\hat{M}$. PME solution, despite its seemingly different ancestry in principle, formal method, and assumptions, agrees with (5.19). This completes our argument.

To maximize the Shannon entropy of M and minimize the Kullback-Leibler divergence between $\hat{M}$ and M, consider the Lagrangian functions:

$$\Lambda(\mu_{ij}, \xi) =$$
$$\sum_{\kappa_{ij}=1} \mu_{ij} \log \mu_{ij} + \sum_{j=1}^{n} \xi_j \left(\beta_j - \sum_{\kappa_{kj}=1} \mu_{kj} \right) +$$
$$\lambda_m \left(x - \sum_{j=1}^{n} \mu_{mj} \right) \tag{5.21}$$

and

$$\hat{\Lambda}(\hat{\mu}_{ij}, \hat{\lambda}) =$$
$$\sum_{\hat{\kappa}_{ij}=1} \hat{\mu}_{ij} \log \frac{\hat{\mu}_{ij}}{\mu_{ij}} + \sum_{i=1}^{m} \hat{\lambda}_i \left(\hat{\alpha}_i - \sum_{\hat{\kappa}_{il}=1} \hat{\mu}_{il} \right). \tag{5.22}$$

For the optimization, we set the partial derivatives to 0, which results in

$$M = rs^\top \circ \kappa \tag{5.23}$$

$$\hat{M} = \hat{r}s^\top \circ \hat{\kappa} \tag{5.24}$$

$$\beta = S\kappa^\top r \tag{5.25}$$

$$\hat{\alpha} = \hat{R}\kappa s \tag{5.26}$$

where $r_i = e^{\zeta_i \lambda_m}, s_j = e^{-1-\xi_j}, \hat{r}_i = e^{-1-\hat{\lambda}_i}$ represent factors arising from the Lagrange multiplier method (ζ was defined in (5.15)). The operator $\circ$ is the entry-wise Hadamard product in linear algebra. $r, s, \hat{r}$ are the vectors containing the $r_i, s_j, \hat{r}_i$, respectively. $R, S, \hat{R}$ are the diagonal matrices with $R_{il} = r_i\delta_{il}, S_{kj} = s_j\delta_{kj}, \hat{R}_{il} = \hat{r}_i\delta_{il}$ (δ is Kronecker delta).

Note that

$$\frac{\beta_j}{\sum_{\hat{\kappa}_{il}=1} \beta_l} = \frac{s_j}{\sum_{\hat{\kappa}_{il}=1} s_l} \text{ for all } (i,j) \in \{1, \ldots, m-1\} \times \{1, \ldots, n\}. \tag{5.27}$$

(5.26) implies

$$\hat{r}_i = \frac{\hat{\alpha}_i}{\sum_{\hat{\kappa}_{il}=1} s_l} \text{ for all } i = 1, \ldots, m-1. \tag{5.28}$$

Consequently,

$$\hat{\beta}_j = s_j \sum_{i=1}^{m-1} \frac{\hat{\kappa}_{ij}\hat{\alpha}_i}{\sum_{\kappa_{il}=1} s_l} \text{ for all } j = 1, \ldots, n. \tag{5.29}$$

(5.29) gives us the same solution as (5.19), taking into account (5.27). Therefore, Wagner conditioning and PME agree.

Wagner-type problems (but not obverse Majerník-type problems) can be solved using JUP and Wagner's ad hoc method. Obverse Majerník-type problems, and therefore all Wagner-type problems, can also be solved using PME and its established and integrated formal method. What at first blush looks like serendipitous coincidence, namely that the two approaches deliver the same result, reveals that JUP is safely incorporated in PME. Not to gain information where such information gain is unwarranted and to process all the available and relevant information is the intuition at the foundation of PME. My results show that this more fundamental intuition generalizes the more specific intuition that ratios of probabilities should remain constant unless they are affected by observation or evidence. Wagner's argument that PME conflicts with JUP is ineffective because it rests on assumptions that proponents of PME naturally reject.

Chapter 6

A Problem for Indeterminate Credal States

6.1 Booleans and Laplaceans

The claim defended in this chapter is that rational agents are subject to a norm requiring sharp credences. I defend this claim in spite of the initially promising features of indeterminate credal states to address problems that sharp credences face in reflecting an agent's doxastic state.

Traditionally, Bayesians have maintained that a rational agent, when holding a credence, holds a sharp credence. It has recently become popular to drop the requirement for credence functions to be sharp. There are now Bayesians who permit a rational agent to hold indeterminate credal states based on incomplete or ambiguous evidence. I will refer to Bayesians who continue to adhere to the classical theory of sharp credences for rational agents as 'Laplaceans' (e.g. Adam Elga and Roger White). I will refer to Bayesians who do not believe that a rational agent's credences need to be sharp as 'Booleans' (e.g. Richard Jeffrey, Peter Walley, Brian Weatherson, and James Joyce). Historically, George Boole and Pierre-Simon Laplace have held positions vaguely related to the ones discussed here (see, for example, Boole, 1854, chapters 16–21).

There is some terminological confusion around the adjectives 'imprecise,' 'indeterminate,' and 'mushy' credences. In the following, I will exclusively refer to indeterminate credences or credal states (abbreviated 'instates') and mean by them a set of sharp credence functions (which some Booleans require to be convex) which it may be rational for an agent to hold within an otherwise orthodox Bayesian framework

(see Jeffrey, 1983).

More formally speaking, let Ω be a set of state descriptions or possible worlds and $\mathcal{X}$ a suitable algebra on Ω. Call $C_{\mathcal{X}}$, which is a set of probability functions on $\mathcal{X}$, a credal state with respect to $\mathcal{X}$. Sometimes the credal state is required to be convex so that $P_1 \in C_{\mathcal{X}}$ and $P_3 \in C_{\mathcal{X}}$ imply $P_2 \in C_{\mathcal{X}}$ if $P_2 = \vartheta P_1 + (1 - \vartheta)P_3$. ϑ is a scalar between 0 and 1, which is multiplied by a probability function using conventional scalar multiplication.

The credal state is potentially different from an agent's doxastic state, which can be characterized in more detail than the credal state (examples will follow). The doxastic state of a rational agent contains all the information necessary to update, infer, and make decisions. Since updating, inference, and decision-making generally needs the quantitative information in a credal state, the credal state is a substate of the doxastic state. Credal states group together doxastic states which are indistinguishable on their formal representation by $C_{\mathcal{X}}$. Laplaceans require that the cardinality of $C_{\mathcal{X}}$ is 1. Booleans have a less rigid requirement of good behaviour for $C_{\mathcal{X}}$, for example they may require it to be a Borel set on the associated vector space if Ω is finite. $C_{\mathcal{X}}$ is a set restricted to probability functions for both Laplaceans and Booleans because both groups are Bayesian.

In the following, I will sometimes say that a credal state with respect to a proposition is a sub-interval of the unit interval, for example $(1/3, 2/3)$. This is a loose way of speaking, since credal states are sets of probability functions, not set-valued functions on a domain of propositions. What I mean, then, is that the credal state identifies $(1/3, 2/3)$ as the range of values that the probability functions in $C_{\mathcal{X}}$ take when they are applied to the proposition in question.

In this chapter, I will introduce a *divide et impera* argument in favour of the Laplacean position. I assume that the appeal of the Boolean position is immediately obvious. Not only is it psychologically implausible that agents who strive for rationality should have all their credences worked out to crystal-clear precision; it seems epistemically doubtful to assign exact credences to propositions about which the agent has little or no information, incomplete or ambiguous evidence (Joyce calls the requirement for sharp credences "psychologically implausible and epistemologically

calamitous," see Joyce, 2005, 156). From the perspective of information theory, it appears that an agent with sharp credences pretends to be in possession of information that she does not have.

The *divide et impera* argument runs like this: I show that the Boolean position is really divided into two mutually incompatible camps, *Bool-A* and *Bool-B*. *Bool-A* has the advantage of owning all the assets of appeal against the Laplacean position. The stock examples brought to bear against sharp credences have simple and compelling instate solutions given *Bool-A*. *Bool-A*, however, has deep conceptual problems which I will describe in detail below.

Bool-B's refinement of *Bool-A* is successful in so far as it resolves the conceptual problems. Their success depends on what I call Augustin's concessions, which undermine all the appeal that the Boolean position as a whole has over the Laplacean position. With a series of examples, I seek to demonstrate that in Simpson Paradox type fashion the Laplacean position looks genuinely inferior to the amalgamated Boolean position, but as soon as the mutually incompatible strands of the Boolean position have been identified, the Laplacean position is independently superior to both.

6.2 Partial Beliefs

Bayesians, whether Booleans or Laplaceans, agree that full belief epistemology gives us an incomplete account of rationality and the epistemic landscape of the human mind. Full belief epistemology is concerned with the acceptance and the rejection of full beliefs, whether an agent may be in possession of knowledge about their contents and what may justify or constitute this knowledge. Bayesians engage in a complementary project investigating partial beliefs. There are belief contents toward which a rational agent has a belief-like attitude characterized by degrees of confidence. These partial beliefs are especially useful in decision theory (for example, betting scenarios). Bayesians have developed a logic of partial beliefs, not dissimilar to traditional logic, which justifies certain partial beliefs in the light of other partial beliefs.

Some epistemologists now seek to reconcile full and partial belief epistemology (see Spohn, 2012; Weatherson, 2012; and Moss, 2013. There is a sense in which, by linking knowledge of chances to its representation in credences, Booleans also seek to reconcile traditional knowledge epistemology concerned with full belief and Bayesian epistemology concerned with partial belief. If the claims in this chapter are correct then the Boolean approach will not contribute to this reconciliation because it mixes full belief and partial belief metaphors in ways that are problematic. The primary task of Bayesians is to explain what partial beliefs are, how they work, and what the norms of rationality are that govern them. The problem for Booleans, as we will see, is that only *Bool-A* has an explanation at hand for how partial beliefs model the epistemic state of the agent in tandem with the way full beliefs do their modeling: as we will see by example, *Bool-A* often looks at partial beliefs as full beliefs about objective chances. *Bool-B* debunks this approach, which leaves the project of reconciliation unresolved.

For the remainder of this section, I want to give the reader a flavour of how appealing the amalgamated version of Booleanism is (the view that a rational agent is permitted to entertain credal states that lack the precision of sharp credences) and then draw the distinction between *Bool-A* and *Bool-B*. Many recently published papers confess allegiance to allowing instates without much awareness of Augustin's and Joyce's refinements in what I call *Bool-B* (for examples see Kaplan, 2010; Hájek and Smithson, 2012; Moss, 2013; Chandler, 2014; and Weisberg, forthcoming). The superiority of the Boolean approach over the Laplacean approach is usually packaged as the superiority of the amalgamated Boolean version, even when the advantages almost exclusively belong to *Bool-A*. Advocates of *Bool-B*, when they defend instates against the Laplacean position, often relapse into *Bool-A*-type argumentation, as we will see by example in a moment.

When we first hear of the advantages of instates, three of them sound particularly persuasive.

- INTERN Instates represent the possibility range for objective chances (objective chances internal to the instate are not believed not to hold, objective chances

external to the instate are believed not to hold).

- INCOMP Instates represent incompleteness or ambiguity of the evidence.
- INFORM Instates are responsive to the information content of evidence.

Here are some examples and explanations. Let a $coin_{\text{x}}$ be a Bernoulli generator that produces successes and failures with probability p_{x} for success, labeled H_{x}, and $1 - p_{\text{x}}$ for failure, labeled T_{x}. Physical coins may serve as Bernoulli generators, if we are willing to set aside that most of them are approximately fair.

> **Example 24: INTERN.** Blake has two Bernoulli generators in her lab, $coin_{\text{i}}$ and $coin_{\text{ii}}$. Blake has a database of $coin_{\text{i}}$ results and concludes on excellent evidence that $coin_{\text{i}}$ is fair. Blake has no evidence about the bias of $coin_{\text{ii}}$. As a Boolean, Blake assumes a sharp credence of $\{0.5\}$ for H_{i} and an indeterminate credal state of $[0, 1]$ for H_{ii}. She feels bad for Logan, her Laplacean colleague, who cannot distinguish between the two cases and who must assign a sharp credence to both H_{i} and H_{ii} (for example, $\{0.5\}$).

> **Example 25: INCOMP.** Blake has another Bernoulli generator, $coin_{\text{iii}}$, in her lab. Her graduate student has submitted $coin_{\text{iii}}$ to countless experiments and emails Blake the resulting bias, but fails to include whether the bias of $2/3$ is in favour of H_{iii} or in favour of T_{iii}. As a Boolean, Blake assumes an indeterminate credal state of $[1/3, 2/3]$ (or $\{1/3, 2/3\}$, depending on the convexity requirement) for H_{iii}. She feels bad for Logan who must assign a sharp credence to H_{iii}. If Logan chooses $\{0.5\}$ as her sharp credence based not unreasonably on symmetry considerations, Logan concurrently knows that her credence gets the bias wrong.

Example 24 also serves as an example for INFORM: one way in which Blake feels bad for Logan is that Logan's $\{0.5\}$ credence for H_{ii} is based on very little information, a fact not reflected in Logan's credence. Walley notes that "the precision of probability models should match the amount of information on which they are based" (Walley, 1991, 34). Joyce explicitly criticizes the information overload for

sharp credences in examples such as example 24. He says about sharp credences of this kind that, despite their maximal entropy compared to other sharp credences, they are "very informative" and "adopting [them] amounts to pretending that you have lots and lots of information that you simply don't have" (Joyce, 2010, 284).

Walley and Joyce appeal to intuition when they promote INFORM. It just feels as if there were more information in a sharp credence than in an instate. Neither of them ever makes this claim more explicit. Joyce admits:

> It is not clear how such an 'imprecise minimum information requirement' might be formulated, but it seems clear that C_1 encodes more information than C_2 whenever $C_1 \subset C_2$, or when C_2 arises from C_1 by conditioning. (Joyce, 2010, 288.)

Since for the rest of the chapter the emphasis will be on INTERN and INCOMP, I will advance my argument against INFORM right away: not only is it not clear how Joyce's imprecise minimum information requirement might be formulated, I see no reason why it should give the results that Joyce envisions. To compare instates and sharp credences informationally, we would need a non-additive set function obeying Shannon's axioms for information (for an excellent summary see Klir, 2006). Attempts for such a generalized Shannon measure have been made, but they are all unsatisfactory. George Klir lists the requirements on page 235 (loc. cit.), and they are worth calling to mind here (a similar list is in Mork, 2013, 363):

(S1) **Probability Consistency** When $\mathcal{D}$ contains only one probability distribution, $\overline{S}$ assumes the form of the Shannon entropy.

(S2) **Set Consistency** When $\mathcal{D}$ consists of the set of all possible probability distributions on $A \subseteq X$, then $\overline{S}(\mathcal{D}) = \log_2 |A|$.

(S3) **Range** The range of $\overline{S}$ is $[0, \log_2 |X|]$ provided that uncertainty is measured in bits.

(S4) **Subadditivity** If $\mathcal{D}$ is an arbitrary convex set of probability distributions on $X \times Y$ and $\mathcal{D}_X$ and $\mathcal{D}_Y$ are the associated sets of marginal probability distributions on X and Y, respectively, then $\overline{S}(\mathcal{D}) \leq \overline{S}(\mathcal{D}_X) + \overline{S}(\mathcal{D}_Y)$.

(S5) **Additivity** If $\mathcal{D}$ is the set of joint probability distributions on $X \times Y$ that is associated with independent marginal sets of probability distributions, $\mathcal{D}_X$ and $\mathcal{D}_Y$, which means that $\mathcal{D}$ is the convex hull of the set $\mathcal{D}_X \otimes \mathcal{D}_Y$, then $\overline{S}(\mathcal{D}) = \overline{S}(\mathcal{D}_X) + \overline{S}(\mathcal{D}_Y)$.

Several things need an explanation here (I have retained Klir's nomenclature). $\mathcal{D}$ is the set of probability distributions constituting the instate. $\overline{S}$ is the proposed generalized Shannon measure defined on the set of possible instates. I will give an example below in (6.2). X is the event space. $\mathcal{D}_X \otimes \mathcal{D}_Y$ is defined as follows:

$$\mathcal{D}_X \otimes \mathcal{D}_Y = \{p(x,y) = p_X(x) \cdot p_Y(y) | x \in X, y \in Y, p_X \in \mathcal{D}_X, p_Y \in \mathcal{D}_Y\}. \quad (6.1)$$

One important requirement not listed is that indeterminateness should give us higher entropy (otherwise Joyce's and Walley's argument will fall flat). Klir's most hopeful contender for a generalized Shannon measure (see his equation 6.61) does not fulfill this requirement:

$$\overline{S}(\mathcal{D}) = \max_{p \in \mathcal{D}} \left\{ -\sum_{x \in X} p(x) \log_2 p(x) \right\}. \quad (6.2)$$

Notice that for any convex instate there is a sharp credence contained in the instate whose generalized Shannon measure according to (6.2) equals the generalized Shannon measure of the instate, but we would expect the entropy of the sharp credence to be lower than the entropy of the instate if it is indeterminate. Jonas Clausen Mork has noticed this problem as well and proposes a modified measure to reflect that more indeterminacy ought to mean higher entropy, all else being equal. He adds the following requirements to Klir's list above (the labels are mine; Mork calls them NC1 and NC2 in Mork, 2013, 363):

(S6) **Weak Monotonicity** If P is a superset of P', then $U(P) \geq U(P')$. A set containing another has at least as great uncertainty value.

(S7) **Strong Monotonicity** If (i) the lower envelope of P is dominated strongly by the lower envelope of P' and (ii) P is a strict superset of P', then $U(P) > U(P')$. When one set strictly contains another with a strictly higher lower bound for at least one hypothesis, the greater set has strictly higher uncertainty value.

I have retained Mork's nomenclature and trust that the reader can see how it lines up with Klir's. Klir's generalized Shannon measure (6.2) fulfills weak monotonicity, but violates strong monotonicity. Mork's proposed alternative is the following (see Mork, 2013, 364):

$$\mathrm{CSU}(\Pi, P) = \max_{p^* \in P} \left\{ -\sum_{i=1}^{n} p^*(h_i) \min_{q^* \in P} \{\log_2 q^*(h_i)\} \right\}. \tag{6.3}$$

Mork fails to establish subadditivity, however, and it is more fundamentally unclear if Klir has not already shown with his disaggregation theory that fulfilling all desiderata (S1)–(S7) is impossible. Some of Klir's remarks seem to suggest this (see, for example, Klir, 2006, 218), but I was unable to discern a full-fledged impossibility theorem in his account. This would be an interesting avenue for further research.

Against the force of INTERN, INCOMP, and INFORM, I maintain that the Laplacean approach of assigning subjective probabilities to partitions of the event space (e.g. objective chances) and then aggregating them by David Lewis' summation formula (see Lewis, 1981, 266f) into a single precise credence function is conceptually tidy and shares many of the formal virtues of Boolean theories. If the bad taste about numerical precision lingers, I will point to philosophical projects in other domains where the concepts we use are sharply bounded, even though our ability to conceive of those sharp boundaries or know them is limited (in particular Timothy Williamson's accounts of vagueness and knowledge). To put it provocatively, this chapter defends a 0.5 sharp credence in heads in all three cases: for a coin of whose bias we are completely ignorant; for a coin whose fairness is supported by a lot of evidence; and even for a coin about whose bias we know that it is either 1/3 or 2/3 for heads.

Statements by Levi and Joyce are representative of how the Boolean position is

most commonly motivated:

> A refusal to make a determinate probability judgment does not derive from a lack of clarity about one's credal state. To the contrary, it may derive from a very clear and cool judgment that on the basis of the available evidence, making a numerically determinate judgment would be unwarranted and arbitrary. (Levi, 1985, 395.)

> As sophisticated Bayesians like Isaac Levi (1980), Richard Jeffrey (1983), Mark Kaplan (1996), have long recognized, the proper response to symmetrically ambiguous or incomplete evidence is not to assign probabilities symmetrically, but to refrain from assigning precise probabilities at all. Indefiniteness in the evidence is reflected not in the values of any single credence function, but in the spread of values across the family of all credence functions that the evidence does not exclude. This is why modern Bayesians represent credal states using sets of credence functions. It is not just that sharp degrees of belief are psychologically unrealistic (though they are). Imprecise credences have a clear epistemological motivation: they are the proper response to unspecific evidence. (Joyce, 2005, 170f.)

Consider therefore the following reasons that incline Booleans to permit instates for rational agents:

(A) The greatest emphasis motivating indeterminacy rests on INTERN, INCOMP, and INFORM.

(B) The preference structure of a rational agent may be incomplete so that representation theorems do not yield single probability measures to represent such incomplete structures.

(C) There are more technical and paper-specific reasons, such as Thomas Augustin's attempt to mediate between the minimax pessimism of objectivists and the Bayesian optimism of subjectivists using interval probability (see Augustin, 2003, 35f); Alan Hájek and Michael Smithson's belief that there may be objectively indeterminate chances in the physical world (see Hájek and Smithson,

2012, 33, but also Hájek, 2003, 278, 307); Jake Chandler's claim that "the sharp model is at odds with a trio of plausible propositions regarding agnosticism" (Chandler, 2014, 4); and Brian Weatherson's claim that for the Boolean position, open-mindedness and modesty may be consistent when for the Laplacean they are not (see Weatherson, 2015, using a result by Gordon Belot, see Belot, 2013).

This chapter mostly addresses (A), while taking (B) seriously as well and pointing towards solutions for it. I am leaving (C) for more specific responses to the issues presented in the cited articles. I will address in section 6.6 Weatherson's more general claim that it is a distinctive and problematic feature of the Laplacean position "that it doesn't really have a good way of representing a state of indecisiveness or open-mindedness" (Weatherson, 2015, 9), i.e. that sharp credences cannot fulfill what I will call the double task (see section 6.6). Weatherson's more particular claim about open-mindedness and modesty is a different story and shall be told elsewhere.

6.3 Two Camps: *Bool-A* and *Bool-B*

My *divide et impera* argument rests on the distinction between two Boolean positions. The difference is best captured by a simple example to show how epistemologists advocate for *Bool-A* or relapse into it, even when they have just advocated the more refined *Bool-B*.

> **Example 26: Skittles.** Every skittles bag contains 42 pieces of candy. It is filled by robots from a giant randomized pile of candies in a warehouse, where the ratio of five colours is 8:8:8:9:9, orange being the last of the five colours. Logan picks one skittle from a bag and tries to guess what colour it is before she looks at it. She has a sharp credence of 9/42 that the skittle is orange.

Bool-A Booleans reject Logan's sharp credence on the basis that she does not know that there are 9 orange skittles in her particular bag. A 9/42 credence suggests to them a knowledge claim on Logan's part, based on very thin evidence, that her

bag contains 9 orange skittles (if this example is not persuasive because the process by which the skittles are chosen is so well known and even a *Bool-A* Booleans should have a sharp credence, the reader is welcome to complicate it to her taste—Bradley and Steele's example is similarly simplistic, see Bradley and Steele, 2016, 6). Logan's doxastic state, however, is much more complicated than her credal state. She knows about the robots and the warehouse. Therefore, her credences that there are k orange skittles in the bag conform to the Bernoulli distribution:

$$C(k) = \binom{42}{k} \left(\frac{9}{42}\right)^{k} \left(\frac{33}{42}\right)^{42-k} \tag{6.4}$$

For instance, her sharp credence that there are in fact 9 orange skittles in the bag is approximately 14.9%. One of Augustin's concessions, the refinements that *Bool-B* makes to *Bool-A*, clarifies that a coherent Boolean position must agree with the Laplacean position that doxastic states are not fully captured by credal states. We will see in the next section why this is the case.

It is a characteristic of *Bool-A*, however, to require that the credal state be sufficient for inference, updating, and decision making. Susanna Rinard, for example, considers it the goal of instates to provide "a complete characterization of one's doxastic attitude" (Rinard, 2015, 5) and reiterates a few pages later that it is "a primary purpose of the set of functions model to represent the totality of the agent's actual doxastic state" (Rinard, 2015, 12).

I will give a few examples of *Bool-A* in the literature, where the authors usually consider themselves to be defending an amalgamated Boolean position which is *pro toto* superior to the Laplacean position. Here is an illustration in Hájek and Smithson.

> **Example 27: Lung Cancer.** Your doctor is your sole source of information about medical matters, and she assigns a credence of $[0.4, 0.6]$ to your getting lung cancer.

Hájek and Smithson go on to say that

> it would be odd, and arguably irrational, for you to assign this proposition a sharper credence—say, 0.5381. How would you defend that assignment? You could say, I don't have to defend it, it just happens to be my credence. But that seems about as unprincipled as looking at your sole source of information about the time, your digital clock, which tells that the time rounded off to the nearest minute is 4:03—and yet believing that the time is in fact 4:03 and 36 seconds. Granted, you may just happen to believe that; the point is that you have no business doing so. (Hájek and Smithson, 2012, 38f.)

This is an argument against Laplaceans by *Bool-A* because it conflates partial belief and full belief. The precise credences in Hájek and Smithson's example, on any reasonable Laplacean interpretation, do not represent full beliefs that the objective chance of getting lung cancer is 0.5381 or that the time of the day is 4:03:36. A sharp credence rejects no hypothesis about objective chances (unlike an instate for *Bool-A*). It often has a subjective probability distribution operating in the background, over which it integrates to yield the sharp credence (it would do likewise in Hájek and Smithson's example for the prognosis of the doctor or the time of the day). The integration proceeds by Lewis' summation formula (see Lewis, 1981, 266f),

$$C(R) = \int_0^1 \zeta P\left(\pi(R) = \zeta\right)\ d\zeta. \tag{6.5}$$

If, for example, S is the proposition that Logan's randomly drawn skittle in example 26 is orange, then

$$C(S) = \sum_{k=0}^{42} \frac{k}{42} \binom{42}{k} \left(\frac{9}{42}\right)^k \left(\frac{33}{42}\right)^{42-k} = 9/42. \tag{6.6}$$

No objective chance $\pi(S)$ needs to be excluded by it. Any updating will merely change the partial beliefs, but no full beliefs. Instates, on the other hand, by giving ranges of acceptable objective chances suggest that there is a full belief that the

objective chance does not lie outside what is indicated by the instate (corresponding to INTERN). When a *Bool-A* advocate learns that an objective chance lies outside her instate, she needs to resort to belief revision rather than updating her partial beliefs. A *Bool-B* advocate can avoid this by accepting one of Augustin's concessions that I will introduce in section 6.5.

Here is another quote revealing the position of *Bool-A*, this time by Kaplan. The example that Kaplan gives is in all relevant respects like example 26, except that he contrasts two cases, one in which the composition of an urn is known and the other where it is not (as the composition of Logan's skittles bag is not known).

> Consider the two cases we considered earlier, and how the difference between them bears on the question as to how confident you should be that (B) the ball drawn will be black. In the first case [where you know the composition of the urn, 100 balls, 50 of which are black], it is clear why you should have a degree of confidence equal to 0.5 that ball drawn from the urn will be black. Your evidence tells you that there is an objective probability of 0.5 that the ball will be black [you only know that there are between 30 and 65 black balls]: it rules every other assignment out either as too low or as too high. In the second case, however, you do not know the objective probability that the ball will be black, because you don't know exactly how many of the balls in the urn are black. Your evidence—thus much inferior in quality to the evidence you have in the first case—doesn't rule out all the assignments your evidence in the first case does. It rules out, as less warranted than the rest, every assignment that gives B a value < 0.3, and every assignment that gives B a value > 0.65. But none of the remaining assignments can reasonably thought to be any more warranted, or less warranted, by your evidence than any other. But then it would seem, at least at first blush, an exercise in unwarranted precision to accede to the requirement, issued by Orthodox Bayesian Probabilism [the Laplacean position], that you choose one of those assignments to be your own. (Kaplan, 2010, 43f.)

While most of these examples have been examples of presenting *Bool-A* as an amalgamated Boolean position, without heed to the refinements of Augustin and

Joyce, it is also the case that refined Booleans belonging to *Bool-B* relapse into *Bool-A* patterns when they argue against the Laplacean position. This is not surprising, because, as we will see, the refined Boolean position *Bool-B* is more coherent than the more simple Boolean position *Bool-A*, but also left without resources to address the problems that have made the Laplacean position vulnerable in the first place.

Joyce, for instance, refers to an example that is again in all relevant respects like example 26 and states that Logan is "committing herself to a definite view about the relative proportions of skittles in the bag" (see Joyce, 2010, 287, pronouns and example-specific nouns changed to fit example 26). Augustin defends the Boolean position with another example of relapse:

> Imprecise probabilities and related concepts ... provide a powerful language which is able to reflect the partial nature of the knowledge suitably and to express the amount of ambiguity adequately. (Augustin, 2003, 34.)

Augustin himself (see section 6.5 on Augustin's concessions) details the demise of the idea that indeterminate credal states can "express the amount of ambiguity adequately." Before I go into these details, however, I need to make the case that Augustin's concessions are necessary in order to refine *Bool-A* and make it more coherent. It is two problems for instates that make this case for us: dilation and the impossibility of learning. Note that these problems are not sufficient to reject instates—they only compel us to refine the more simple Boolean position via Augustin's concessions. The final game is between *Bool-B* and Laplaceans, where I will argue that the *Bool-B* position has lost the intuitive appeal of the amalgamated Boolean position to present solutions to prima facie problems facing the Laplacean position.

6.4 Dilation and Learning

Here are two potential problems for Booleans:

- DILATION Instates are vulnerable to dilation.

- OBTUSE Instates do not permit learning.

Both of these can be resolved by making Augustin's concessions. I will introduce these problems in the present section 6.4, then Augustin's concessions in the next section 6.5, and the implications for the more general disagreement between Booleans and Laplaceans in section 6.6.

6.4.1 Dilation

Consider the following example for DILATION (see White, 2010, 175f and Joyce, 2010, 296f).

> **Example 28: Dilation.** Logan has two Bernoulli generators, $coin_{\text{iv}}$ and $coin_{\text{v}}$. She has excellent evidence that $coin_{\text{iv}}$ is fair and no evidence about the bias of $coin_{\text{v}}$. Logan's graduate student independently tosses both $coin_{\text{iv}}$ and $coin_{\text{v}}$. Then she tells Logan whether the results of the two tosses correspond or not ($H_{\text{iv}} \equiv H_{\text{v}}$ or $H_{\text{iv}} \equiv T_{\text{v}}$, where $X \equiv Y$ means $(X \wedge Y) \vee (\neg X \wedge \neg Y)$). Logan, who has a sharp credence for H_{v}, takes this information in stride, but she feels bad for Blake, whose credence in H_{iv} dilates to $[0, 1]$ even though Blake shares Logan's excellent evidence that $coin_{\text{iv}}$ is fair.

Here is why Blake's credence in H_{iv} must dilate. Her credence in H_{v} is $[0, 1]$, by stipulation. Let $c(X)$ be the range of probabilities represented by Blake's instate with respect to the proposition X, for example $c(H_{\text{v}}) = [0, 1]$. Then

$$c(H_{\text{iv}} \equiv H_{\text{v}}) = c(H_{\text{iv}} \equiv T_{\text{v}}) = \{0.5\} \tag{6.7}$$

because the tosses are independent and $c(H_{\text{iv}}) = \{0.5\}$ by stipulation. Next,

$$c(H_{\text{iv}}|H_{\text{iv}} \equiv H_{\text{v}}) = c(H_{\text{v}}|H_{\text{iv}} \equiv H_{\text{v}}) \tag{6.8}$$

where $c(X|Y)$ is the updated instate after finding out Y. Booleans accept (6.8) because they are Bayesians and update by standard conditioning. Therefore,

$$c(H_{\text{iv}}|H_{\text{iv}} \equiv H_{\text{v}}) = c(H_{\text{v}}|H_{\text{iv}} \equiv H_{\text{v}}) = c(H_{\text{v}}) = [0, 1]. \tag{6.9}$$

To see that (6.9) is true, note that in the rigorous definition of a credal state as a set of probability functions, each probability function P in Blake's instate for the proposition $H_{\text{v}}|H_{\text{iv}} \equiv H_{\text{v}}$ has the following property by Bayes' theorem:

$$P(H_{\text{v}}|H_{\text{iv}} \equiv H_{\text{v}}) = \frac{P(H_{\text{iv}})P(H_{\text{v}})}{P(H_{\text{iv}})P(H_{\text{v}}) + P(T_{\text{iv}})P(T_{\text{v}})} \tag{6.10}$$

In our loose way of speaking of credal states as set-valued functions of propositions, Blake's updated instate for H_{iv} has dilated from $\{0.5\}$ to $[0, 1]$ (for another example of dilation with discussion see Bradley and Steele, 2016).

This does not sound like a knock-down argument against Booleans (it is investigated in detail in Seidenfeld and Wasserman, 1993), but Roger White uses it to derive implications from instates which are worrisome.

> **Example 29: Chocolates.** Four out of five chocolates in the box have cherry fillings, while the rest have caramel. Picking one at random, what should my credence be that it is cherry-filled? Everyone, including the staunchest [Booleans], seems to agree on the answer 4/5. Now of course the chocolate I've chosen has many other features, for example this one is circular with a swirl on top. Noticing such features could hardly make a difference to my reasonable credence that it is cherry filled (unless of course I have some information regarding the relation between chocolate shapes and fillings). Often chocolate fillings do correlate with their shapes, but I haven't the faintest clue how they do in this case or any reason to suppose they correlate one way rather than another ... the further result is that while my credence that the chosen chocolate is cherry-filled should be 4/5 prior to viewing it, once I see its shape (whatever shape it happens to be) my credence that it is cherry-filled should dilate to become [indeterminate]. But this is just not the way we think about such matters. (White, 2010, 183.)

I will characterize the problems that dilation causes for the Boolean position by three aspects (all of which originate in White, 2010):

- RETENTION This is the problem in example 29. When we tie the outcome of a random process whose objective chance we know to the outcome of a random process whose chance we do not know, White maintains that we should be entitled to *retain* the credence that is warranted by the known objective chance. One worry about the Boolean position in this context is that credences become excessively dependent on the mode of representation of a problem (Howson calls this the description-relativity worry).

- REPETITION Consider again example 28, although this time Blake runs the experiment 10,000 times. Each time, her graduate student tells her whether $H^n_{\text{iv}} \equiv H^n_{\text{v}}$ or $H^n_{\text{iv}} \equiv T^n_{\text{v}}$, n signifying the n-th experiment. After running the experiment that many times, approximately half of the outcomes of $coin_{\text{iv}}$ are heads. Now Blake runs the experiment one more time. Again, the Boolean position mandates dilation, but should Blake not just on inductive grounds persist in a sharp credence of 0.5 for H_{iv}, given that about half of the coin flips so far have come up heads? Attentive readers will notice a sleight of hand here: as many times as Blake performs the experiment, it must not have any evidential impact on the assumptions of example 28, especially that the two coin flips remain independent and that the credal state for $coin_{\text{v}}$ is still, even after all these experiments, maximally indeterminate.

- REFLECTION Consider again example 28. Blake's graduate student will tell Blake either $H_{\text{iv}} \equiv H_{\text{v}}$ or $H_{\text{iv}} \equiv T_{\text{v}}$. No matter what the graduate student tells Blake, Blake's credence in H_{iv} dilates to an instate of $[0, 1]$. Blake therefore is subject to Bas van Fraassen's reflection principle stating that

$$P^a_t(A|p^a_{t+x}(A) = r) = r \tag{6.11}$$

  where P^a_t is the agent a's credence function at time t, x is any non-negative number, and $p^a_{t+x}(A) = r$ is the proposition that at time $t + x$, the agent a will

bestow degree r of credence on the proposition A (see van Fraassen, 1984, 244). Van Fraassen had sharp credences in mind, but it is not immediately obvious why the reflection principle should not also hold for instates.

To address the force of RETENTION, REPETITION, and REFLECTION, Joyce hammers home what I have called Augustin's concessions. According to Joyce, none of White's attacks succeed if a refinement of *Bool-A*'s position takes place and instates are not required either to reflect knowledge of objective chances or doxastic states. I will address this in detail in section 6.5, concluding that Augustin's concessions are only necessary based on REFLECTION, whereas solutions for RETENTION and REPETITION have less far-reaching consequences and do not impugn the Boolean position of *Bool-A*.

6.4.2 Learning

Here is an example for OBTUSE (see Rinard's objection cited in White, 2010, 84 and addressed in Joyce, 2010, 290f). It presumes Joyce's supervaluationist semantics of instates (see Hájek, 2003; Joyce, 2010, 288; and Rinard, 2015; for a problem with supervaluationist semantics see Lewis, 1993; and an alternative which may solve the problem see Weatherson, 2015, 7), for which Joyce uses the helpful metaphor of committee members, each of whom holds a sharp credence. The instate consists then of the set of sharp credences from each committee member: for the purposes of updating, for example, each committee member updates as if she were holding a sharp credence. The aggregate of the committee members' updated sharp credences forms the updated instate. Supervaluationist semantics also permits comparisons, when for example a partial belief in X is stronger than a partial belief in Y because all committee members have sharp credences in X which exceed all the sharp credences held by committee members with respect to Y.

Example 30: Learning. Blake has a Bernoulli generator in her lab, $coin_{\text{vi}}$, of whose bias she knows nothing and which she submits to experiments. At first, Blake's instate for H_{vi} is $(0, 1)$. After a few experiments, it looks like $coin_{\text{vi}}$ is fair. However,

> as committee members crowd into the centre and update their sharp credences to something closer to 0.5, they are replaced by extremists on the fringes. The instate remains at $(0, 1)$.

It is time now to examine refinements of the Boolean position to address these problems.

6.5 Augustin's Concessions

Joyce has defended instates against DILATION and OBTUSE, making Augustin's concessions (AC1) and (AC2). I am naming them after Thomas Augustin, who has some priority over Joyce in the matter. Augustin's concessions distinguish *Bool-A* and *Bool-B*, the former of which does not make the concessions and identifies partial beliefs with full beliefs about objective chances. A sophisticated view of partial beliefs recognizes that they are sui generis, which necessitates a substantial reconciliation project between full belief epistemology and partial belief epistemology. My task at hand is to agree with the refined position of *Bool-B* in their argument against *Bool-A* that this reconciliation project is indeed substantial and that partial beliefs are not full beliefs about objective chances; but also that *Bool-B* fails to summon arguments against the Laplacean position without relapsing into an unrefined version of indeterminacy.

Here, then, are Augustin's concessions:

(AC1) Credal states do not adequately represent doxastic states. The same instate can reflect different doxastic states, even when the difference in the doxastic states matters for updating, inference, and decision making.

(AC2) Instates do not represent full belief claims about objective chances. White's *Chance Grounding Thesis* is not an appropriate characterization of the Boolean position.

I agree with Joyce that (AC1) and (AC2) are both necessary and sufficient to resolve DILATION and OBTUSE for instates. I disagree with Joyce about what this means for

an overall recommendation to accept the Boolean rather than the Laplacean position. After I have already cast doubt on INFORM, I will show that (AC1) and (AC2) neutralize INTERN and INCOMP, the major impulses for rejecting the Laplacean position.

Indeterminacy imposes a double task on credences (representing both uncertainty and available evidence) that they cannot coherently fulfill. I will present several examples where this double task stretches instates to the limits of plausibility. Joyce's idea that credences can represent balance, weight, and specificity of the evidence (in Joyce, 2005) is inconsistent with the use of indeterminacy. Joyce himself, in response to DILATION and OBTUSE, gives the argument why this is the case (see Joyce, 2010, 290ff, for OBTUSE; and Joyce, 2010, 296ff, for DILATION). Let us begin by looking more closely at how (AC1) and (AC2) protect *Bool-B* from DILATION and OBTUSE.

6.5.1 Augustin's Concession (AC1)

(AC1) says that credences do not adequately represent a doxastic state. The same instate can reflect different doxastic states, where the difference is relevant to updating, inference, and decision making.

Augustin recognizes the problem of inadequate representation before Joyce, with specific reference to instates: "The imprecise posterior does no longer contain all the relevant information to produce optimal decisions. Inference and decision do not coincide any more" (Augustin, 2003, 41) (see also an example for inadequate representation of evidence by instates in Bradley and Steele, 2014, 1300). Joyce rejects the notion that identical instates encode identical beliefs by giving two examples. The first one is problematic. The second one, which is example 28 given earlier, addresses the issue of DILATION more directly. Here is the first example.

> **Example 31: Three-Sided Die.** Suppose $\mathcal{C}'$ and $\mathcal{C}''$ are defined on a partition $\{X, Y, Z\}$ corresponding to the result of a roll of a three sided-die. Let $\mathcal{C}'$ contain all credence functions defined on $\{X, Y, Z\}$ such that $c(Z) \geq 1/2$, and let $\mathcal{C}''$ be the subset of $\mathcal{C}''$ whose members also satisfy $c(X) = c(Y)$ (see Joyce, 2010, 294).

Joyce then goes on to say,

> It is easy to show that $\mathcal{C}'$ and $\mathcal{C}''$ generate the same range of probabilities for all Boolean combinations of $\{X, Y, Z\}$ ... but they are surely different: the $\mathcal{C}''$-person believes everything the $\mathcal{C}'$-person believes, but she also regards X and Y as equiprobable.

Example 31 is problematic because $\mathcal{C}'$ and $\mathcal{C}''$ do not generate the same range of probabilities: if, as Joyce says, $c(Z) \geq 1/2$, then $c(X) = c(Y)$ implies $c(X) \leq 1/4$ for $\mathcal{C}''$, but not for $\mathcal{C}'$. What Joyce wants to say is that the same instate can encode doxastic states which are relevantly different when it comes to updating probabilities, and the best example for this is example 28 itself.

To explain this in more detail, we need to review for a moment what Lewis means by the Principal Principle and by inadmissibility. The Principal Principle requires that my knowledge of objective chances is reflected in my credence, unless there is inadmissible evidence. Inadmissible evidence would for instance be knowledge of a coin toss outcome, in which case of course I do not need to have a credence for it corresponding to the bias of the coin. In example 28, I could use the Principal Principle in the spirit of RETENTION to derive a contradiction to the Boolean formalism.

$$H_{\text{iv}} \equiv H_{\text{v}} \text{ does not give anything away about } H_{\text{iv}}, \tag{6.12}$$

therefore

$$c(H_{\text{iv}}|H_{\text{iv}} \equiv H_{\text{v}}) = c(H_{\text{iv}}) \tag{6.13}$$

by the Principal Principle and in contradiction to (6.9).

Joyce explains how (6.12) is false and blocks the conclusion (6.13), which would undermine the Boolean position. $H_{\text{iv}} \equiv H_{\text{v}}$ is clearly inadmissible, even without (AC1), since it is information that not only changes Blake's doxastic state, but also her credal state.

We would usually expect more information to sharpen our credal states (see Walley's anti-dilation principle and his response to this problem in 1991, 207 and

299), an intuition violated by both DILATION and OBTUSE. As far as DILATION is concerned, however, the loss of precision is in principle not any more surprising than information that increases the Shannon entropy of a sharp credence.

> **Example 32: Rumour.** A rumour that the Canadian prime minister has been assassinated raises your initially very low probability that this event is taking place today to approximately 50%.

It is true for both sharp and indeterminate credences that information can make us less certain about things. This is the simple solution for RETENTION. If one of Joyce's committee members has a sharp credence of 1 in H_{v} and learns $H_{\text{iv}} \equiv H_{\text{v}}$, then her sharp credence for H_{iv} should obviously be 1 as well; ditto and mutatis mutandis for the committee member who has a sharp credence of 0 in H_{v}. (AC1) is unnecessary.

Here is how dilation is as unproblematic as a gain in entropy after more information in example 32:

> **Example 33: Dilating Urns.** You are about to draw a ball from an urn with 200 balls (100 red, 100 yellow). Just before you draw, you receive the information that the urn has two chambers which are obscured to you as you draw the ball, one with 99 red balls and 1 yellow ball, the other with 1 red ball and 99 yellow balls.

Dilation from a sharp credence of $\{0.5\}$ to an instate of $[0.01, 0.99]$ (or $\{0.01, 0.99\}$, depending on whether convexity is required) is unproblematic, although the example prefigures that there is something odd about the Boolean conceptual approach. The example licences a 99:1 bet for one of the colours (if the instate is interpreted as upper and lower previsions), which inductively would be a foolish move. This is a problem that arises out of the Boolean position quite apart from DILATION and in conjunction with betting licences, which we will address in example 36.

So far we have not found convincing reasons to accept (AC1). Neither dilation as a phenomenon nor Lewis' inadmissibility criterion need to compel a *Bool-A* advocate to admit (AC1), despite Joyce's claims in the abstract of his paper that reactions to

dilation "are based on an overly narrow conception of imprecise belief states which assumes that we know everything there is to know about a person's doxastic attitudes once we have identified the spreads of values for her imprecise credences." Not only does example 31 not give us the desired results, we can also resolve RETENTION and REPETITION without recourse to (AC1). It is, in the final analysis, REFLECTION which will make the case for (AC1).

It is odd that Joyce explicitly says that the REPETITION argument "goes awry by assuming that credal states which assign the same range of credences for an event reflect the same opinions about that event" (Joyce, 2010, 304), when in the following he makes an air-tight case against the anti-Boolean force of REPETITION without ever referring to (AC1). Joyce's case against REPETITION is based on the sleight of hand to which I made reference when I introduced REPETITION. There is no need to repeat Joyce's argument here.

Let me add that despite my agreement with Joyce on how to handle REPETITION, although we disagree that it has anything to do with (AC1), a worry about the Boolean position with respect to REPETITION lingers. Joyce requires that Blake, in so far as Blake wants to be rational, has a maximally indeterminate credal state for H_{iv}^{10000} after hearing from her graduate student. The oddity of this, when we have just had about 5000 heads and 5000 tails, remains. Blake's maximally indeterminate credal state rests on her stubborn conviction that her credal state must be inclusive of the objective chance of $coin_{\text{iv}}^{10000}$ landing heads.

Logan, if she were to do this experiment, would relax, reason inductively as well as based on her prior probability, and give H_{iv} a credence of 0.5, since her credal state is not in the same straight-jacket of underlying objective chances—just as in example 26 Logan is able to have a sharp credence of 9/42 for picking an orange skittle, when her credence that there are 9 orange skittles in the bag is only 14.9%. Logan's attitude may serve as additional motivation for Augustin's second concession (AC2). A proponent of *Bool-B* would rather stand with Logan and hold that the relationship between instates and objective chances is much looser than for *Bool-A*, but more about this in subsection 6.5.2.

We are left with REFLECTION, which bears most of the burden in Joyce's ar-

gument for (AC1). It is indeed odd that a rational agent should have two strictly distinct credal states C_1 and C_2, when C_2 follows upon C_1 from a piece of information that says either X or $\neg X$—but it does not matter which of the two. Why does the rational agent not assume C_2 straight away? Joyce introduces an important distinction for van Fraassen's reflection principle: It is not the degree of credence that is decisive as in van Fraassen's original formulation of the principle, but the doxastic state. X and $\neg X$ (in example 28, they refer to $H_{\text{iv}} \equiv H_{\text{v}}$ and $H_{\text{iv}} \equiv T_{\text{v}}$) both lead from a sharp credence of 0.5 to an instate of $[0, 1]$ for H_{iv}, but this updated instate reflects different doxastic states. In Joyce's words, "the beliefs about H_{iv} you will come to have upon learning $H_{\text{iv}} \equiv H_{\text{v}}$ are complementary to the beliefs you will have upon learning $H_{\text{iv}} \equiv T_{\text{v}}$" (Joyce, 2010, 304). Committee members representing complementary beliefs agree on the instate, but single committee members take opposite views depending on the information, X or $\neg X$ (besides Joyce, see also Topey, 2012, 483). Credal states keep track only of the committee's aggregate credal state, whereas doxastic states keep track of each committee member's individual sharp credences. For Joyce, this is a purely formal distinction, not to be confused with questions about the believer's psychology (see Joyce, 2010, 288).

This resolves the anti-Boolean force of REFLECTION by making the concession (AC1). Ironically, of course, not being able to represent a doxastic state, but only to reflect it inadequately, was just the problem that Laplacean sharp credences had which Boolean instates were supposed to fix. One way *Bool-B* could respond is like this:

> **Riposte:** If you are going to make a difference between *representing* a doxastic state and *reflecting* a doxastic state, where the former poses an identity relationship between credal states and doxastic states and the latter a supervenience relationship, then all you have succeeded in making me concede is that both instates and sharp credences reflect doxastic states. Instates may still be more successful in reflecting doxastic states than sharp credences are and so solve the abductive task of explaining partial beliefs better.

There are several reasons why I doubt this line of argument is successful. The

Laplacean position has formal advantages, for example that it is able to cooperate with information theory, an ability which *Bool-B* lacks. There is no coherent theory of how relations between instates can be evaluated on the basis of information and entropy, which are powerful tools when it comes to justifying fundamental Bayesian tenets (see Shore and Johnson, 1980; Giffin, 2008). Furthermore, the Laplacean position is conceptually tidy. It distinguishes between the quantifiable aspects of a doxastic state, which it integrates to yield the sharp credence, and other aspects of the evidence, such as incompleteness or ambiguity. Instates dabble in what I will call the double task: trying to reflect both aspects without convincing success in either.

6.5.2 Augustin's Concession (AC2)

(AC2) says that instates do not reflect knowledge claims about objective chances. White's *Chance Grounding Thesis* (which White does not endorse, being a Laplacean) is not an appropriate characterization of the Boolean position.

> **Chance Grounding Thesis (CGT):** Only on the basis of known chances can one legitimately have sharp credences. Otherwise one's spread of credence should cover the range of possible chance hypotheses left open by your evidence. (White, 2010, 174)

Joyce considers (AC2) to be as necessary for a coherent Boolean view of partial beliefs, blocking OBTUSE, as (AC1) is, blocking DILATION (see Joyce, 2010, 289f).

OBTUSE is related to VACUITY, another problem for Booleans:

- VACUITY If one were to be committed to the principle of regularity, that all states of the world considered possible have positive probability (for a defence see Edwards et al., 1963, 211); and to the solution of Henry Kyburg's lottery paradox, that what is rationally accepted should have probability 1 (for a defence of this principle see Douven and Williamson, 2006); and the CGT, that one's spread of credence should cover the range of possible chance hypotheses left open by the evidence (implied by much of Boolean literature); then one's instate would always be vacuous.

Booleans must deny at least one of the premises to avoid the conclusion. Joyce denies the CGT, giving us (AC2). It is by no means necessary to sign on to regularity and to the above-mentioned solution of Henry Kyburg's lottery paradox in order to see how (AC2) is a necessary refinement of *Bool-A*. The link between objective chances and credal states expressed in the CGT is suspect for many other reasons. I have referred to them passim, but will not go into more detail here.

6.6 The Double Task

Sharp credences have a single task: to reflect epistemic uncertainty as a tool for updating, inference, and decision making. They cannot fulfill this task without continued reference to the evidence which operates in the background. To use an analogy, credences are not sufficient statistics with respect to updating, inference, and decision making. What is remarkable about Joyce's response to DILATION and OBTUSE is that Joyce recognizes that instates are not sufficient statistics either. But this means that they fail at the double task which has been imposed on them: to represent both epistemic uncertainty and relevant features of the evidence.

In the following, I will provide a few examples where it becomes clear that instates have difficulty representing uncertainty because they are tangled in a double task which they cannot fulfill.

> **Example 34: Aggregating Expert Opinion.** Blake has no information whether it will rain tomorrow (R) or not except the predictions of two weather forecasters. One of them forecasts 0.3 on channel GPY, the other 0.6 on channel QCT. Blake considers the QCT forecaster to be significantly more reliable, based on past experience.

An instate corresponding to this situation may be $[0.3, 0.6]$ (see Walley, 1991, 214), but it will have a difficult time representing the difference in reliability of the experts. We could try $[0.2, 0.8]$ (since the greater reliability of QCT suggests that the chance of rain tomorrow is higher rather than lower) or $[0.1, 0.7]$ (since the greater reliability of QCT suggests that its estimate is more precise), but it remains obscure what the criteria are.

A sharp credence of $P(R) = 0.53$, for example, does the right thing. Such a credence says nothing about any beliefs that the objective chance is restricted to a subset of the unit interval, but it accurately reflects the degree of uncertainty that the rational agent has over the various possibilities. Beliefs about objective chances make little sense in many situations where we have credences, since it is doubtful even in the case of rain tomorrow that there is an urn of nature from which balls are drawn. What is really at play is a complex interaction between epistemic states (for example, experts evaluating meteorological data) and the evidence which influences them.

As we will see in the next example, it is an advantage of sharp credences that they do not exclude objective chances, even extreme ones, because they express partial belief and do not suggest, as instates do for *Bool-A*, that there is full belief knowledge that the objective chance is a member of a proper subset of the possibilities (for an example of a crude version of indeterminacy that reduces partial beliefs to full beliefs see Levi, 1981, 540, "inference derives credal probability from knowledge of the chances of possible outcomes"; or Kaplan, 2010, 45, "you should rule out all and only the assignments the evidence warrants your regarding as too high or too low, and you should remain in suspense between those degree of confidence assignments that are left").

> **Example 35: Precise Credences.** Logan's credence for rain tomorrow, based on the expert opinion of channel GPY and channel QCT (she has no other information) is 0.53. Is it reasonable for Logan, considering how little evidence she has, to reject the belief that the chance of rain tomorrow is 0.52 or 0.54; or to prefer a 52.9 cent bet on rain to a 47.1 cent bet on no rain?

The first question in example 35 is as confused as the *Bool-A* confusion found in example 27 discussed earlier. As for the second question in example 35: why would we prefer a 52.9 cent bet on rain to a 47.1 cent bet on no rain, given that we do not possess the power of discrimination between these two bets? If X and Y are two propositions, then it is important not to confuse the claim that it is reasonable to hold both X and Y with the claim that it is reasonable to hold either X (without

Y) or Y (without X). It is the reasonableness of holding X and Y concurrently that is controversial, not the reasonableness of holding Y (without holding X) when it is reasonable to hold X.

Let $U(S, Z, t)$ mean "it is rational for S to believe Z at time t." Then U is exportable (see Rinard, 2015, 6) if and only if $U(S, X, t)$ and $U(S, Y, t)$ imply $U(S, X \wedge Y, t)$. Beliefs somehow grounded in subjectivity, such as beliefs about etiquette or colour perception are counter-examples for the exportability of U. Vagueness also gives us cases of non-exportability. Rinard considers the connection between vagueness and indeterminacy to be an argument in favour of indeterminacy.

My argument is that non-exportability blunts an argument against the Laplacean position. In a moment, I will talk about anti-luminosity, the fact that a rational agent may not be able to distinguish psychologically between a 54.9 cent bet on an event and a 45.1 bet on its negation, when her sharp credence is 0.55. She must reject one of them not to incur sure loss, so proponents of indeterminacy suggest that she choose one of them freely without being constrained by her credal state or reject both of them. I claim that a sharp credence will make a recommendation between the two so that only one of the bets is rational given her particular credence, but that does not mean that another sharp credence which would give a different recommendation may not also be rational for her to have. Partial beliefs are non-exportable.

The answer to the second question in example 35 ties in with the issue of incomplete preference structure referred to above as motivation (B) for instates (see page 136).

> It hardly seems a requirement of rationality that belief be precise (and preferences complete); surely imprecise belief (and corresponding incomplete preferences) are at least rationally permissible. (Bradley and Steele, 2014, 1288, for a similar sentiment see Kaplan, 2010, 44.)

The development of representation theorems beginning with Frank Ramsey (followed by increasingly more compelling representation theorems in Savage, 1954; and Jeffrey, 1965; and numerous other variants in contemporary literature) bases probability and utility functions of an agent on her preferences, not the other way around.

Once completeness as an axiom for the preferences of an agent is jettisoned, indeterminacy follows automatically. Indeterminacy may thus be a natural consequence of the proper way to think about credences in terms of the preferences that they represent.

In response, preferences may very well logically and psychologically precede an agent's probability and utility functions, but that does not mean that we cannot inform the axioms we use for a rational agent's preferences by undesirable consequences downstream. Completeness may sound like an unreasonable imposition at the outset, but if incompleteness has unwelcome consequences for credences downstream, it is not illegitimate to revisit the issue.

Timothy Williamson goes through this exercise with vague concepts, showing that all upstream logical solutions to the problem fail and that it has to be solved downstream with an epistemic solution (see Williamson, 1996). Vague concepts, like sharp credences, are sharply bounded, but not in a way that is luminous to the agent (for anti-luminosity see chapter 4 in Williamson, 2000). Anti-luminosity answers the second question in example 35: the rational agent prefers the 52.9 cent bet on rain to a 47.1 cent bet on no rain based on her sharp credence without being in a position to have this preference necessarily or have it based on physical or psychological ability (for the analogous claim about knowledge see Williamson, 2000, 95).

In a way, advocates of indeterminacy have solved this problem for us. There is strong agreement among most of them that the issue of determinacy for credences is not an issue of elicitation (sometimes the term 'indeterminacy' is used instead of 'imprecision' to underline this difference; see Levi, 1985, 395). The appeal of preferences is that we can elicit them more easily than assessments of probability and utility functions. The indeterminacy issue has been raised to the probability level (or moved downstream) by indeterminacy advocates themselves who feel justifiably uncomfortable with an interpretation of their theory in behaviourist terms. So it shall be solved there, and this chapter makes an appeal to reject indeterminacy on this level. The solution then has to be carried upstream (or lowered to the logically more basic level of preferences), where we recognize that completeness for preferences is after all a desirable axiom for rationality and "perfectly rational agents always have

perfectly sharp probabilities" (Elga, 2010, 1). When Levi talks about indeterminacy, it also proceeds from the level of probability judgment to preferences, not the other way around (see Levi, 1981, 533).

Note also that Booleans customarily talk about imprecise probabilities, but seldom about imprecise utility. Bradley and Steele "do not anticipate any problems in extending the results to the case of both imprecise probabilities and utilities" (Bradley and Steele, 2016, 9.), but without having investigated this any further I suspect that there are counterintuitive implications downstream for imprecise utility as well.

> **Example 36: Monkey-Filled Urns.** Let urn A contain 4 balls, two red and two yellow. A monkey randomly fills urn B from urn A with two balls. We draw from urn B (a precursor to this example is in Jaynes and Bretthorst, 2003, 160).

One reasonable sharp credence of drawing a red ball is 0.5, following Lewis' summation formula for the different combinations of balls in urn B and symmetry considerations. This solution is more intuitive in terms of further inference, decision making, and betting behaviour than a credal state of $\{0, 1/2, 1\}$ or $[0, 1]$ (depending on the convexity requirement), since this instate would licence an exorbitant bet in favour of one colour, for example one that costs \$9,999 and pays \$10,000 if red is drawn and nothing if yellow is drawn.

How a bet is licenced is different on various Boolean accounts. Rinard, for example, contrasts a moderate account with a liberal account (see Rinard, 2015, 7). According to the liberal account, the \$9,999 bet is licenced, whereas according to the moderate account, it is only indeterminate whether the bet is licenced. The moderate account does not take away from the force of example 36, where it should be determinate that a \$9,999 bet is not licenced.

To make example 36 more vivid consider a Hand Urn, where you draw by hand from an urn with 100 balls, 50 red balls and 50 yellow balls. When your hand retreats from the urn, does it not contain either a red ball or a yellow ball and so serve itself as an urn, from which in a sense you draw a ball? Your hand contains one ball, either red or yellow, and the indeterminate credal state that it is one or the other should

be $[0, 1]$. This contradicts our intuition that our credence should be a sharp 0.5 and raises again the mode of representation issue, as in RETENTION. Sharp credences are more resistent to partition variance when the partition is along lines that appear irrelevant to the question (such as symbols on pieces of chocolate as in example 29 or how a ball is retrieved from an urn as in example 36).

> **Example 37: Three Prisoners.** Prisoner X_1 knows that two out of three prisoners (X_1, X_2, X_3) will be executed and one of them pardoned. She asks the warden of the prison to tell her the name of another prisoner who will be executed, hoping to gain knowledge about her own fate. When the warden tells her that X_3 will be executed, X_1 erroneously updates her probability of pardon from $1/3$ to $1/2$, since either X_1 or X_2 will be spared.

Walley maintains that for the Monty Hall problem and the Three Prisoners problem, the probabilities of a rational agent should dilate rather than settle on the commonly accepted solutions. For the Three Prisoners problem, there is a compelling case for standard conditioning and the result that the credence for prisoner X_1 to receive a pardon ought not to change after the update (see section 7.4). Walley's dilated solution would give prisoner X_1 hope on the doubtful possibility (and unfounded assumption) that the warden might prefer to provide X_3's (rather than X_2's) name in case prisoner X_1 was pardoned.

This example brings an interesting issue to the forefront. Sharp credences often reflect independence of variables where such independence is unwarranted. Booleans (more specifically, detractors of the principle of indifference or the principle of maximum entropy, principles which are used to generate sharp credences for rational agents) tend to point this out gleefully. They prefer to dilate over the possible dependence relationships, independence included. Dilation in example 28 is an instance of this. The fallacy in the argument for instates, illustrated by the Three Prisoners problem, is that the probabilistic independence of sharp credences does not imply independence of variables. Only the converse is correct. Probabilistic independence may simply reflect an averaging process over various dependence relationships, independence again included.

In the Three Prisoners problem, there is no evidence about the degree or the direction of the dependence, and so prisoner X_1 should take no comfort in the information that she receives. The rational prisoner's probabilities will reflect probabilistic independence, but make no claims about causal independence. Walley has unkind things to say about sharp credences and their ability to respond to evidence (for example that their "inferences rarely conform to evidence", see Walley, 1991, 396), but in this case it appears to me that they outperform the Boolean approach.

Joyce also commits himself to dilation for the Three Prisoners problem and would thus have to call the conventional solution defective epistemology (see Joyce, 2010, 292), for in the comparable case of example 28 he states that "behaving as if these two events are independent amounts to pulling a statistical correlation out of thin air!" (Joyce, 2010, 300.) The two events in question are H_{iv} and $H_{\text{iv}} \equiv H_{\text{v}}$ in example 28, but they could just as well be 'X_1 will be executed' and 'warden says X_3 will be executed' in example 37.

To conclude, a Boolean in the light of Joyce's two Augustinian concessions has three alternatives, of which I favour the third: (a) to find fault with Joyce's reasoning as he makes those concessions; (b) to think (as Joyce presumably does) that the concessions are compatible with the promises of Booleans, such as INTERN, INCOMP, and INFORM, to solve prima facie problems of sharp credences; or (c) to abandon the Boolean position because (AC1), (AC2), and an array of examples in which sharp credences are conceptually and pragmatically more appealing show that the initial promise of the Boolean position is not fulfilled.

Chapter 7

Judy Benjamin

7.1 Goldie Hawn and a Test Case for Full Employment

Probability kinematics is the field of inquiry asking how we should update a probability distribution in the light of evidence. If the evidence comes as an event, it is relatively uncontroversial to use conditional probabilities (standard conditioning). Sometimes, however, the evidence may not relate the certainty of an event but a reassessment of its uncertainty or its probabilistic relation to other events (see Jeffrey, 1965, 153ff), expressible in a shift in expectation (see Hobson, 1971). Jeffrey conditionalization can deal with some of these cases, but not with all of them (see figure 7.1). Bas van Fraassen has come up with an example for a case in which we cannot apply Jeffrey conditionalization. The example is from the 1980 comedy film *Private Benjamin* (see van Fraassen, 1981), in which Goldie Hawn portrays a Jewish-American woman (Judy Benjamin) who joins the U.S. Army.

> **Example 38: Judy Benjamin.** In van Fraassen's interpretation of the movie, Judy Benjamin is on an assignment and lands in a place where she is not sure of her location. She is on team Blue. Because of the map, initially the probability of being in Blue territory equals the probability of being in Red territory, and the probability of being in the Red Second Company area equals the probability of being in the Red Headquarters area. Her commanders then inform Judy by radio that in case she is in Red territory, her chance of being in the Red Headquarters area is three times the chance of being in the Red Second Company area.

The question is what Judy's appropriate response is to this new evidence. We

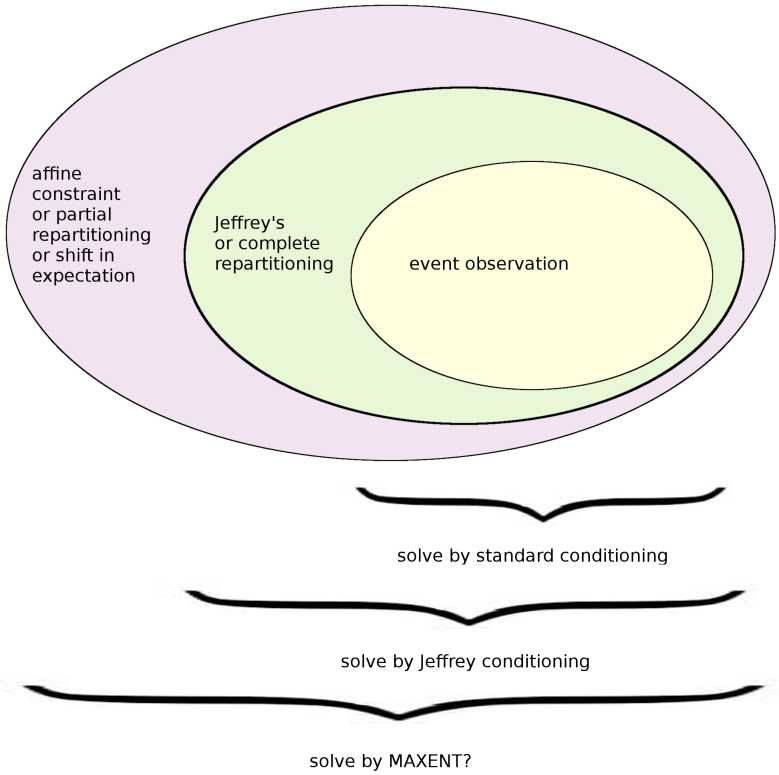

Figure 7.1: Information that leads to unique solutions for probability updating using PME must come in the form of an affine constraint (a constraint in the form of an informationally closed and convex space of probability distributions consistent with the information). All information that can be processed by Jeffrey conditioning comes in the form of an affine constraint, and all information that can be processed by standard conditioning can also be processed by Jeffrey conditioning. The solutions of PME are consistent with the solutions of Jeffrey conditioning and standard conditioning where the latter two are applicable.

cannot apply standard conditioning, because there is no immediately obvious event space in which we can condition on an event of which we are certain. Grove and Halpern (1997) have offered a proposal for constructing such event spaces and then conditioning on the event that Judy Benjamin receives the information that she receives from her commanders. They admit, however, that the construction of such spaces (sometimes called retrospective conditioning) is an exercise in filling in missing

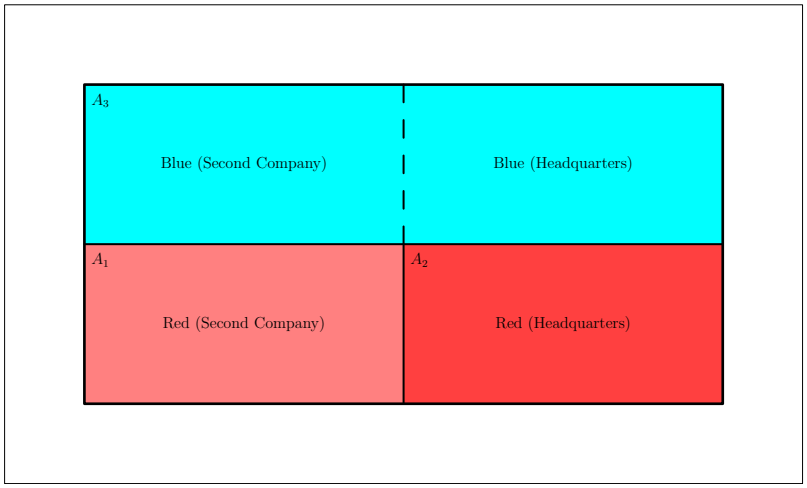

Figure 7.2: Judy Benjamin's map. Blue territory (A_3) is friendly and does not need to be divided into a Headquarters and a Second Company area.

details and supplying information not contained in the original problem.

If we assume that the attempt fails to define an event on which Judy Benjamin could condition her probabilities, we are left with two possibilities. Her new information (it is three times as likely to land in A_2 than as to land in A_1, see figure 7.2 and the details of the problem in the next section) may mean that we have a redistribution of a complete partition of the probabilities. This is called Jeffrey conditioning and calls for Jeffrey's rule. Jeffrey's rule is contested in some circles, but we will for this project accept its validity in probability kinematics. We will see in what follows that some make the case that Jeffrey conditioning is the correct way to solve the Judy Benjamin problem. For reasons provided in the body of the chapter their case is implausible.

The third possibility to solve this problem (after standard conditioning and Jeffrey conditioning) is to consult a highly contested updating procedure: the principle of maximum entropy (PME for short). PME can be applied to any situation in which we have a completely quantified probability distribution and an affine constraint (we will explain the nature of affine constraints in more detail later). If our new evidence is the observation of an event (or simply certainty about an event that we did not

have previously), then the event provides an affine constraint and can be used for updating by means of standard conditioning. If our new evidence is a redistribution of probabilities where we can apply Jeffrey's rule, then the redistribution provides an affine constraint and can be used for updating by means of Jeffrey's rule. These two possibilities, however, do not exhaust affine constraints. The Judy Benjamin problem illustrates the third possibility where the affine constraint only redistributes some groups of probabilities and leaves open the question how this will affect the probabilities not included in this redistribution.

Advocates of PME claim that in this case the probabilities should be adjusted so that they are minimally affected (we make this precise by using information theory) while at the same time according with the constraint. Opponents of this view grant that PME is an important tool of probability kinematics. Noting that some results of PME are difficult to accept (such as in the Judy Benjamin case), however, they urge us to embrace a more pluralistic, situation-specific methodology.

Joseph Halpern, for example, writes in *Reasoning About Uncertainty* that "there is no escaping the need to understand the details of the application" (Halpern, 2003, 423) and concludes that PME is a valuable tool, but should be used with care (see Grove and Halpern, 1997, 110), explicitly basing his remark on the counterintuitive behaviour of the Judy Benjamin problem. Diaconis and Zabell state: "any claims to the effect that maximum-entropy revision is the only correct route to probability revision should be viewed with considerable caution" (Diaconis and Zabell, 1982, 829). "Great caution" (1994, 456) is also what Colin Howson and Allan Franklin advise about the more basic claim that the updated probabilities provided by PME are as like the original probabilities as it is possible to be given the constraints imposed by the data.

In the same vein, Igor Douven and Jan-Willem Romeijn agree with Richard Bradley that "even Bayes' rule 'should not be thought of as a universal and mechanical rule of updating, but as a technique to be applied in the right circumstances, as a tool in what Jeffrey terms *the art of judgment*.' In the same way, determining and adapting the weights ... may be an art, or a skill, rather than a matter of calculation or derivation from more fundamental epistemic principles" (Douven and

Romeijn, 2009, 16) (for the Bradley quote see Bradley, 2005, 362).

What is lacking in the literature is a response by PME advocates to the counterintuitive behaviour of the cases repeatedly quoted by their adversaries. This is especially surprising as we are not dealing with an array of counter-examples but only a handful, the Judy Benjamin problem being prime among them. In Halpern's textbook, for example, the reasoning is as follows: PME is a promising candidate which delivers unique updated probability distributions; but, unfortunately, there is counterintuitive behaviour in one specific case, the Judy Benjamin case (see Halpern, 2003, 110, 119); therefore, we must abide by the eclectic principle of considering not only PME, but also lower and upper probabilities, Dempster-Shafer belief functions, possibility measures, ranking functions, relative likelihoods, and so forth. The human inquirer is the final arbiter between these conditionalization methods.

At the heart of our investigation are two incompatible but independently plausible intuitions regarding Judy's choice of updated probabilities for her location. We will undermine the notion that PME's solution for the Judy Benjamin problem is counterintuitive. The intuition that PME's solution for the Judy Benjamin problem violates (call it T1) is based on fallacious independence and uniformity assumptions. There is another powerful intuition (call it T2) that conflicts with T1 and obeys PME. Therefore, Halpern does not give us sufficient grounds for the eclecticism advocated throughout his book. We will show that another intuitive approach, the powerset approach, lends significant support to the solution provided by PME for the Judy Benjamin problem, especially in comparison to intuition T1, many of whose independence and uniformity assumptions it shares.

We have no proof that PME is the only rationally defensible objective method to update probabilities given an affine constraint. The literature outlines many of the 'nice properties' of PME. It seamlessly generalizes standard conditioning and Jeffrey's rule where they are applicable (see Caticha and Giffin, 2006). It underlies the entropy concentration phenomenon described in Jaynes' standard work *Probability Theory: the Logic of Science*, which contains other arguments in favour of PME (some of which you may recognize by family resemblance in the rest of this chapter). Entropy concentration refers to the unique property of PME solution to have other

distributions which obey the affine constraint cluster around it. Shore and Johnston have shown that under certain rationality assumptions PME provides the unique solution to problems of probability updating (see Shore and Johnson, 1980). When used to make predictions whose quality is measured by a logarithmic score functions, posterior probabilities provided by PME result in minimax optimal decisions (see Topsøe, 1979, Walley, 1991, and Grünwald, 2000). Under a logarithmic scoring rule these posterior probabilities are in some sense optimal.

Despite all these nice properties, we want the reader to follow us in a more simple line of argument. When new evidence is provided to us, it is rational to adjust our beliefs minimally in light of it. We do not want to draw out more information from the new evidence than necessary. We are the first to admit that there are numerous problems here that need addressing. What do we mean by rationality? What are the semantics of the word 'minimal'? What are the formal properties of such posterior probabilities? Are they unique? Are they compatible with other intuitive methods of updating? Are there counter-intuitive examples that would encourage us to give up on this line of thought rather than live with its consequences? Given some decent answers to these questions, however, we feel that PME cuts a good figure as a first pass to provide objective solutions to these types of problems, and the burden on opponents who usually deny that there are such objective solutions exist grows heavy.

The distinctive contribution of this chapter is to show why the reasoning of the opponents of PME in the Judy Benjamin case is flawed. They make independence assumptions that on closer inspection do not hold up. We provide a number of scenarios consistent with the information in the problem which violate these independence assumptions. That does not mean that the information given in the problem suggests these scenarios, it only means that we are not entitled to make those independence assumptions. That, in turn, does not privilege PME solution, although PME does not lean on independence assumptions that other solutions illegitimately make. PME, however, confronts us with a much stronger claim than merely providing a passable or useful solution to the Judy Benjamin problem: it claims to much greater generality and, to use a term abjured by many formal epistemologists, to objectivity. These claims must be motivated elsewhere, and the nature of their normativity is a matter

of debate (for a pragmatic approach see Caticha, 2012). We are only showing that opponents cannot claim an easy victory by pulling out old Judy Benjamin.

There is a long-standing disagreement between (mostly) philosophers on the one hand and (mostly) physicists on the other hand. The philosophers claim that updating probabilities is irreducibly accompanied by thoughtful deliberation with the choice between different updating procedures depending on individual problems. The physicists claim that problems are ill-posed if they do not contain the information necessary to let a non-arbitrary, objective procedure (such as PME) arrive at a unique updated probability distribution. In the literature, Judy Benjamin serves as an example widely taken to count in favour of the philosophers. It is taken to support what I call the full employment theorem of probability kinematics.

7.2 Two Intuitions

There are two pieces of information relevant to Judy Benjamin when she decides on her updated probability assignment. We will call them (MAP) and (HDQ). As in figure 7.2, A_1 is the Red Second Company area, A_2 is the Red Headquarters area, A_3 is Blue territory. Judy presumably wants to be in Blue territory, but if she is in Red territory, she would prefer their Second Company area (where enemy soldiers are not as well-trained as in the Headquarters area).

(MAP) Judy has no idea where she is. Because of the map, her probability of being in Blue territory equals the probability of being in Red territory, and the probability of being in the Red Second Company area equals the probability of being in the Red Headquarters area.

(HDQ) Her commanders inform Judy that in case she is in Red territory, her chance of being in their Headquarters area is three times the chance of being in their Second Company area.

In formal terms (sloppily writing A_i for the event of Judy being in A_i),

$$2 \cdot P(A_1) = 2 \cdot P(A_2) = P(A_3) \tag{MAP}$$

$$\vartheta = P(A_2|A_1 \cup A_2) = \frac{3}{4} \tag{HDQ}$$

(HDQ) is partial information because in contrast to the kind of evidence we are used to in Bayes' formula (such as 'an even number was rolled'), and to the kind of evidence needed for Jeffrey's rule (where a partition of the whole event space and its probability redistribution is required, not only $A_1 \cup A_2$, but see here the objections in Douven and Romeijn, 2009), the scenario suggests that Bayesian conditionalization and Jeffrey's rule are inapplicable. We are interested in the most defensible updated probability assignment(s) and will express them in the form of a normalized odds vector (q_1, q_2, q_3), following van Fraassen (1981). q_i is the updated probability $Q(A_i)$ that Judy Benjamin is in A_i. Let P be the probability distribution prior to the new observation and p_i the individual 'prior' probabilities. These probabilities are not to be confused with prior probabilities that precede any kind of information. In the spirit of probability update, or probability kinematics, we will for the rest of the article refer to prior probabilities as probabilities prior to an observation and the subsequent update. The q_i sum to 1 (this differs from van Fraassen's canonical odds vector, which is proportional to the normalized odds vector but has 1 as its first element). We define

$$t = \frac{\vartheta}{1 - \vartheta} \tag{7.1}$$

t is the factor by which (HDQ) indicates that Judy's chance of being in A_2 is greater than being in A_1. In Judy's particular case, $t = 3$ and $\vartheta = 0.75$. Two intuitions guide the way people think about Judy Benjamin's situation.

T1 (HDQ) does not refer to Blue territory and should not affect $P(A_3)$: $q_3 = p_3 (= 0.50)$.

There is another, conflicting intuition (due to Peter Williams via personal communication with van Fraassen, see van Fraassen, 1981, 379):

T2 If the value of ϑ approaches 1 (in other words, t approaches infinity) then q_3 should approach 2/3 as the problem reduces to one of ordinary conditioning. (HDQ) would turn into 'if you are in Red territory you are almost certainly in the Red Headquarters area.' Considering (MAP), q_3 should approach 2/3.

Continuity considerations pose a contradiction to T1. (These considerations are strong enough that Luc Bovens uses them as an assumption to solve Adam Elga's Sleeping Beauty problem by parity of reasoning in Bovens, 2010.) To parse these conflicting intuitions, we will introduce several methods to provide G, the function that maps ϑ to the appropriate normalized updated odds vector (q_1, q_2, q_3).

The first method is extremely simple and accords with intuition T1: $G_{\text{ind}}(\vartheta) = (0.5(1-\vartheta), 0.5\vartheta, 0.5)$. In Judy's particular case with $t = 3$ the normalized odds vector is (ind stands for independent):

$$G_{\text{ind}}(0.75) = (0.125, 0.375, 0.500) \tag{7.2}$$

Both Grove and Halpern (1997) and Douven and Romeijn (2009) make a case for this distribution. Grove and Halpern use standard conditioning on the event of the message being transmitted to Judy. Douven and Romeijn use Jeffrey's rule (because they believe that T1 is in this case so strong that $Q(A_3) = P(A_3)$ is as much of a constraint as (MAP) and (HDQ), yielding a Jeffrey partition).

T1, however, conflicts with the symmetry requirements outlined in van Fraassen et. al. (1986). Van Fraassen introduces various updating methods which do not conflict with those symmetry requirements, the most notable of which is PME. Shore and Johnson have already shown that, given certain assumptions (which have been

heavily criticized, e.g. in Uffink, 1996), PME produces the unique updated probability assignment according with these assumptions. The minimum information discrimination theorem of Kullback and Leibler (see for example Csiszár, 1967, section 3) demonstrates how Shannon's entropy and the Kullback-Leibler Divergence formula can provide the least informative updated probability assignment (with reference to the prior probability assignment) obeying the constraint posed by the evidence. The idea is to define a space of probability distributions, make sure that the constraint identifies a closed, convex subset in this space, and then determine which of the distributions in the closed, convex subset is least distant from the prior probability distribution in terms of information (using the minimum information discrimination theorem). It is necessary for the uniqueness of this least distant distribution that the subset be closed and convex (in other words, that the constraint be affine, see Csiszár, 1967).

For Judy Benjamin, PME suggests the following normalized odds vector:

$$G_{\max}(0.75) \approx (0.117, 0.350, 0.533) \tag{7.3}$$

The updated probability of being on Blue territory (A_3) has increased from 50% to approximately 53%. Grove and Halpern find this result "highly counterintuitive" (Grove and Halpern, 1997, 2). Van Fraassen summarizes the worry:

> It is hard not to speculate that the dangerous implications of being in the enemy's Headquarters area are causing Judy Benjamin to indulge in wishful thinking, her indulgence becoming stronger as her conditional estimate of the danger increases. (van Fraassen, 1981, 379.)

There are two ways in which we can arrive at result (7.3). We may use the constraint rule for cross-entropy derived in subsection 3.2.3 or use the Kullback-Leibler Divergence and differentiate it to obtain where it is minimal. The constraint rule has the advantage of providing results when the derivative of the Kullback-Leibler Divergence is difficult to find. This not being the case for Judy, we go the easier route

of the second method here and provide a more general justification for the constraint rule in subsection 3.2.3, together with its application to the Judy Benjamin case.

The Kullback-Leibler Divergence is

$$D(Q,P) = \sum_{i=1}^{m} q_i \log_2 \frac{q_i}{p_i}. \tag{7.4}$$

We fill in the explicit details from Judy Benjamin's situation and differentiate the expression to obtain the minimum (by setting the derivative to 0).

$$\frac{\partial}{\partial q_1}(q_1 \log_2(4q_1) + tq_1 \log_2(4tq_1) + (1-(t+1)q_1)\log_2 2(1-(t+1)q_1)) = 0 \tag{7.5}$$

The resulting expression for $G_{\max}$ is

$$G_{\max}(\vartheta) = \left(\frac{C}{1+Ct+C}, t\frac{C}{1+Ct+C}, 1-(t+1)\frac{C}{1+Ct+C}\right), \tag{7.6}$$

where

$$C = 2^{-\frac{t\log_2 t+t+1}{1+t}}. \tag{7.7}$$

Figures 7.3 and 7.4 show in diagram form the distribution of (q_1, q_2, q_3) depending on the value of ϑ (between 0 and 1), respectively following intuition T1 and PME. Notice that in accordance with intuition T2, PME provides a result where $q_3 \to 2/3$ for ϑ approaching 0 or 1.

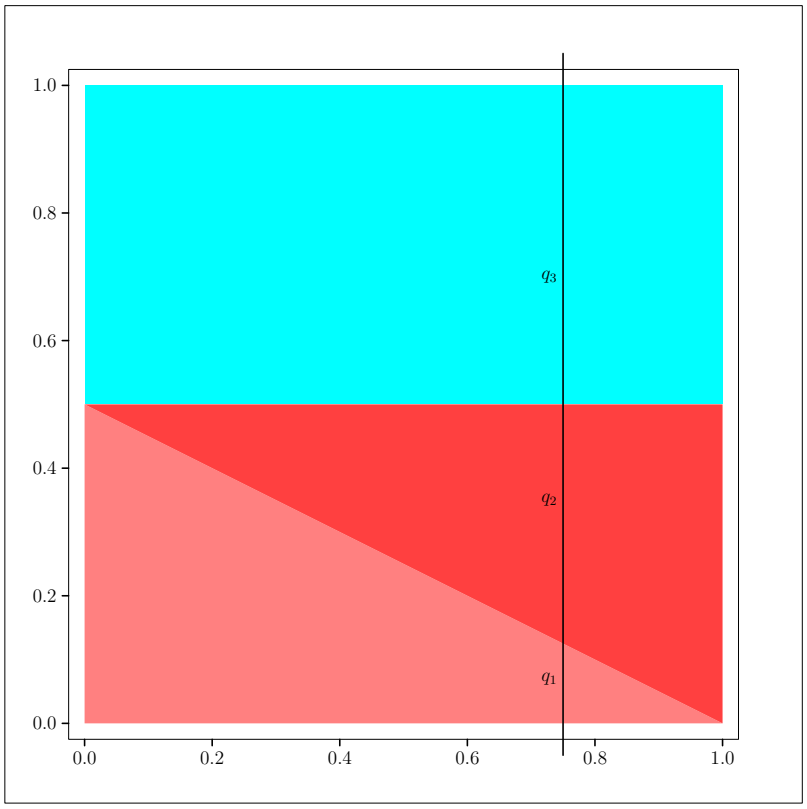

Figure 7.3: Judy Benjamin's updated probability assignment according to intuition T1. $0 < \vartheta < 1$ forms the horizontal axis, the vertical axis shows the updated probability distribution (or the normalized odds vector) (q_1, q_2, q_3). The vertical line at $\vartheta = 0.75$ shows the specific updated probability distribution $G_{\text{ind}}(0.75)$ for the Judy Benjamin problem.

7.3 Epistemic Entrenchment

Consider two future events A and B. You have partial belief in whether they will occur and assign probabilities to them. Then you learn that A entails B. How does this information affect your probability assignment for event A? If A is causally independent of B then your updated probability for it should equal the original probability.

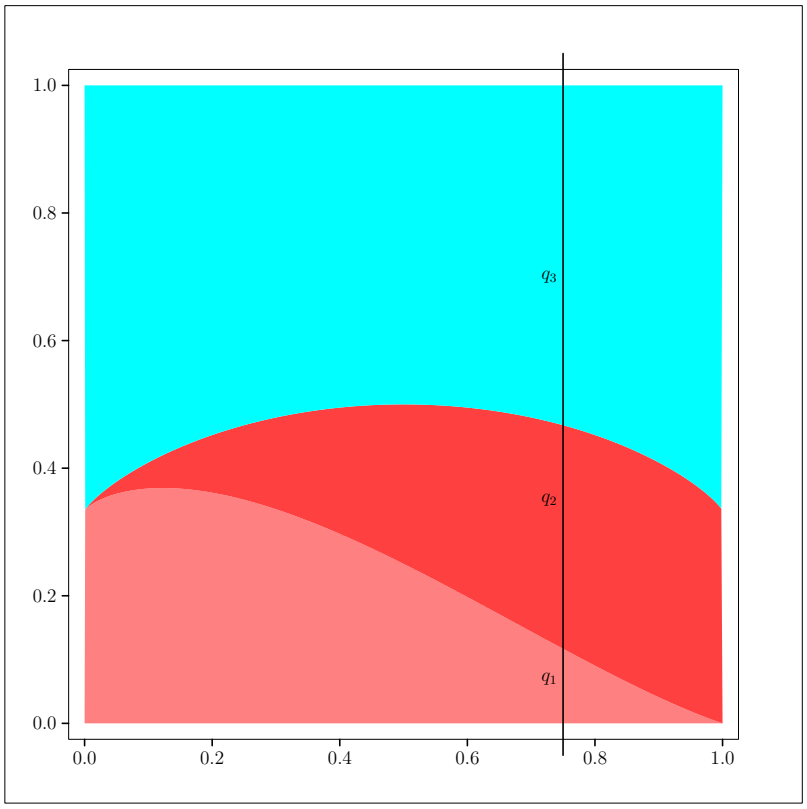

Figure 7.4: Judy Benjamin's updated probability assignment using PME. $0 < \vartheta < 1$ forms the horizontal axis, the vertical axis shows the updated probability distribution (or the normalized odds vector) (q_1, q_2, q_3). The vertical line at $\vartheta = 0.75$ shows the specific updated probability distribution $G_{\max}(0.75)$ for the Judy Benjamin problem.

Example 39: Sundowners at the Westcliff. Whether Sarah and Marian have sundowners at the Westcliff hotel tomorrow may initially not depend on the weather at all, but if they learn that there will be a wedding and the hotel's indoor facilities will be closed to the public, then rainfall tomorrow implies no sundowners.

Learning this conditional does not affect the probability of the antecedent (rainfall tomorrow), because the antecedent is causally independent of the consequent.

Example 40: Heist. A jeweler has been robbed, and Kate has reason to assume that Henry might be the robber. Kate knows that he is not capable of actually injuring

another person, but he may very well engage in robbery. When Kate hears from the investigator that the robber also shot the jeweler she concludes that Henry is not the robber. (A similar example is in Leendert Huisman's paper "Learning from Simple Indicative Conditionals," to be published in *Erkenntnis*, involving an exam and a skiing vacation.)

Kate has learned a conditional and adjusted the probability for the antecedent. The reason for this is that Kate was epistemically entrenched to uphold her belief in Henry's nonviolent nature. Updating probabilities upon learning a conditional depends on epistemic entrenchments. Examples 39 and 40 are from Douven and Romeijn, 2009).

In Judy Benjamin's case, (HDQ) is also a conditional. If Judy is in Red territory, she is more likely to be in the Headquarters area. According to PME, the updated probability for the antecedent of this conditional is raised. It appears that PME preempts our epistemic entrenchments and nolens volens assigns a certain degree of confirmation to the antecedent of a learned conditional. This degree of confirmation depends on the causal dependency of the antecedent on the consequent. In this section we will focus on the independence assumptions that are improperly imported into the Judy Benjamin case by detractors of PME. We will be particularly critical of Douven and Romeijn, who hold that the Judy Benjamin case is a case for Adam's conditioning, where the antecedent is left alone in the updating of probabilities.

Even though T1 is an understandably strong intuition, it does not take into account that the information given to Judy by her commanders may indeed be dependent on whether she is in Blue or in Red territory. To underline this objection to intuition T1 consider three scenarios, any of which may form the basis of the partial information provided by her commanders.

I Judy is dropped off by a pilot who flips two coins. If the first coin lands H, then Judy is dropped off in Blue territory, otherwise in Red territory. If the second coin lands H, she is dropped off in the Headquarters area, otherwise in the Second Company area. Judy's commanders find out that the second coin is biased $\vartheta : 1 - \vartheta$ toward H with $\vartheta = 0.75$. The normalized odds vector is

$G_{\text{I}}(0.75) = (0.125, 0.375, 0.500)$ and agrees with T1, because the choice of Blue or Red is completely independent from the choice of the Red Headquarters area or the Red Second Company area.

II The pilot randomly lands in any of the four quadrants and rolls a die. If she rolls an even number, she drops off Judy. If not, she takes her to another (or the same, the choice happens with replacement) randomly selected quadrant to repeat the procedure. Judy's commanders find out, however, that for A_1, the pilot requires a six to drop off Judy, not just an even number. The normalized odds vector in this scenario is $G_{\text{II}}(0.75) = (0.1, 0.3, 0.6)$ and does not accord with T1.

III Judy's commanders have divided the map into 24 congruent rectangles, A_3 into twelve, and A_1 and A_2 into six rectangles each (see figures 7.5 and 7.6). They have information that the only subsets of the 24 rectangles in which Judy Benjamin may be located are such that they contain three times as many A_2 rectangles than A_1 rectangles. The normalized odds vector in this scenario is $G_{\text{III}}(0.75) \approx (.108, .324, .568)$ (evaluating almost 17 million subsets).

I–III demonstrate the contrast between scenarios when independence is true and when it is not. Douven and Romeijn's capital mistake in their paper is that they assume that the Judy Benjamin problem is analogous to example 39. Sarah, however, knows that whether it rains or not is independent of her activity the next night, whereas in Judy Benjamin we have no evidence of such independence, as scenario II makes clear. This is not to say that scenario II is the scenario that pertains in Judy Benjamin's case. It only says that there is no natural assumption in Judy Benjamin's case that the probabilities are independent of each other in light of the new evidence, for scenario II is perfectly natural (whether it is true or not is a completely different question) and reveals how dependence is consistent with the information that Judy Benjamin receives.

Douven and Romeijn's strong independence claim relying on intuition T1 leads them to apply Jeffrey's rule to the Judy Benjamin problem with the additional

constraint $q_3 = p_3$. They claim that in most cases "the learning of a conditional is or would be irrelevant to one's degree of belief for the conditional's antecedent ... the learning of the relevant conditional should intuitively leave the probability of the antecedent unaltered" (Douven and Romeijn, 2009, 9).

This, according to Douven and Romeijn, is the usual epistemic entrenchment and applies in full force to the Judy Benjamin problem. They give an example where the epistemic entrenchment could go the other way and leave the consequent rather than the antecedent unaltered (Kate and Henry, see Douven and Romeijn, 2009, 13). The idea of epistemic entrenchment is at odds with PME and seems to imply just what the full employment theorem claims: judgments so framed "will depend on the judgmental skills of the agent, typically acquired not in the inductive logic class but by subject specific training" (Bradley, 2005, 349). To pursue the relation between the semantics of conditionals, PME, and the full employment theorem would take us too far afield at present and shall be undertaken elsewhere. For the Judy Benjamin problem, it is not clear why Douven and Romeijn think that the way the problem is posed implies a strong epistemic entrenchment for Adams conditioning (Adams conditioning is the kind of conditioning that will leave the antecedent alone). Scenarios II-III provide realistic alternatives where Adams conditioning is inappropriate.

Judy Benjamin may also receive (HDQ) because her informers have found out that Red Headquarters troops have occupied the entire Blue territory ($q_1 = 3p_1, q_2 = p_2, q_3 = 0$, the epistemic entrenchment is with respect to q_2); because they have found out that Blue troops have occupied two-thirds of the Red Second Company area ($q_1 = p_1, q_2 = (1/3)p_2, q_3 = (4/3)p_3$, the epistemic entrenchment is with respect to q_1); or because they have found out that Red Headquarters troops have taken over half of the Red Second Company area ($q_1 = (1/2)p_1, q_2 = (3/2)p_2, q_3 = p_3$, the epistemic entrenchment is with respect to q_3 and what Douven and Romeijn take to be an assumption in the wording of the problem). There is nothing in the problem that supports Douven and Romeijn's narrowing of the options. The table reiterates these options, with the last line representing intuition (T1) and the epistemic entrenchment defended by Douven and Romeijn.

Epistemic entrenchment	q_1	q_2	q_3
with respect to A_1	1/4	3/4	0
with respect to A_2	1/12	1/4	2/3
with respect to A_3	1/8	3/8	1/2

7.4 Coarsening at Random

Another at first blush forceful argument that PME's solution for the Judy Benjamin problem is counterintuitive has to do with coarsening at random, or CAR for short. CAR involves using more naive (or coarse) event spaces in order to arrive at solutions to probability updating problems. The CAR condition requires that conditioning on these coarse events does not give a result that is inconsistent with conditioning on a more finely-grained event that has been observed. The mechanics are spelled out in (2003). Grünwald and Halpern see a parallel between the Judy Benjamin problem and Martin Gardner's *Three Prisoners* problem, see example 37 on page 158.

According to Grünwald and Halpern, for problems of this kind (Judy Benjamin, Three Prisoners, Monty Hall) there are naive and sophisticated spaces to which we can apply probability updates. If A uses the naive space, for example, he comes to the following conclusion: of the three possibilities that (A,B), (A,C), or (B,C) are executed, the warden's information excludes (A,C). (A,B) and (B,C) are left over, and because A has no information about which one of these is true his chance of not being executed is 0.5. His chance of survival has increased from one third to one half.

Grünwald and Halpern show, correctly, that the application of the naive space is illegitimate because the CAR condition does not hold. More generally, Grünwald and Halpern show that updating on the naive space rather than the sophisticated space is legitimate for event type observations always when the set of observations is pairwise disjoint or, when the events are arbitrary, only when the CAR condition holds. For Jeffrey type observations, there is a generalized CAR condition which applies likewise. For affine constraints on which we cannot use Jeffrey conditioning (or, a fortiori, standard conditioning) MAXENT "essentially never gives the right

results" (Grünwald and Halpern, 2003, 243).

Grünwald and Halpern conclude that "working with the naive space, while an attractive approach, is likely to give highly misleading answers" (246), especially in the application of PME to naive spaces as in the Judy Benjamin case "where applying [PME] leads to paradoxical, highly counterintuitive results" (245). For the Three Prisoners problem, Jaynes' constraint rule would supposedly proceed as follows: the vector of prior probabilities for (A,B), (A,C), and (B,C) is $(1/3, 1/3, 1/3)$. The constraint is that the probability of (A,C) is zero, and a simple application of the constraint rule yields $(1/2, 0, 1/2)$ for the vector of updated probabilities. The CAR condition for the naive space does not hold, therefore the result is misleading.

By analogy, using the constraint rule on the naive space for the Judy Benjamin problem yields $(0.117, 0.350, 0.533)$, but as the CAR condition fails in even the simplest settings for affine constraints ("CAR is (roughly speaking) guaranteed *not* to hold except in 'degenerate' situations" (251), emphasis in the original), it certainly fails for the Judy Benjamin problem, for which constructing a sophisticated space is complicated (see Grove and Halpern, 1997, where the authors attempt such a construction by retrospective conditioning).

The analogy, however, is misguided. The constraint rule has been formally shown to generalize Jeffrey conditioning, which in turn has been shown to generalize standard conditioning (the authors admit as much in Grünwald and Halpern, 2003, 262). We can solve both the Monty Hall problem and the Three Prisoners problem by standard conditioning, not using the naive space, but simply using the correct space for the probability update. For the Three Prisoners problem, for example, the warden will say either 'B' or 'C' in response to A's inquiry. Because A has no information that would privilege either answer the probability that the warden says 'B' and the probability that the warden says 'C' equal each other and therefore equal 0.5. Here is the difference between using the naive space and using the correct space, but either way using standard conditional probabilities:

$$\begin{array}{l} P(\text{`A is pardoned'}|\text{`B will be executed'}) \\ \frac{P(\text{`A is pardoned'})}{P(\text{`A is pardoned'})+P(\text{`C is pardoned'})} \end{array} \begin{array}{l} = \\ = \frac{1}{2} \text{ (incorrect)} \end{array} \tag{7.8}$$

$$\begin{array}{l} P(\text{`A is pardoned'}|\text{`warden says B will be executed'}) \\ \frac{P(\text{`A is pardoned' and `warden says B will be executed'})}{P(\text{`warden says B will be executed'})} \end{array} \begin{array}{l} = \\ = \frac{1/6}{1/2} = \frac{1}{3} \text{ (correct)} \end{array} \tag{7.9}$$

Why is the first equation incorrect and the second one correct? Information theory gives us the right answer: in the first equation, we are conditioning on a watered down version of the evidence (watered down in a way that distorts the probabilities because we are not 'coarsening at random'). 'Warden says B will be executed' is sufficient but not necessary for 'B will be executed.' The former proposition is more informative than the latter proposition (its probability is lower). Conditioning on the latter proposition leaves out relevant information contained in the wording of the problem.

Because PME always agrees with standard conditioning, PME gives the correct result for the Three Prisoners problem. For the Judy Benjamin problem, there is no defensible sophisticated space and no watering down of the evidence in what Grünwald and Halpern call the 'naive' space. The analogy between the Three Prisoners problem and the Judy Benjamin problem as it is set up by Grünwald and Halpern fails because of this crucial difference. A successful criticism would be directed at the construction of the 'naive' space: this is what we just accomplished for the Three Prisoners problem. There is no parallel procedure for the Judy Benjamin problem. The 'naive' space is all we have, and PME is the appropriate tool to deal with this lack of information.

7.5 The Powerset Approach

In this section, we will focus on scenario III and consider what happens when the grain of the partition becomes finer. We call this the powerset approach. Two remarks are in order. On the one hand, the powerset approach has little independent appeal. The reason behind using PME is that we want our evidence to have just the right influence on our updated probabilities, i.e. neither over-inform nor under-inform. There is no corresponding reason why we should update our probabilities using the powerset approach. On the other hand, what the powerset approach does is lend support to another approach. In this task, it is persuasive because it tells us what would happen if we were to divide the event space into infinitesimally small, uniformly weighed, and independent 'atomic' bits of information.

In the process of arriving at the formal result, the powerset approach resembles an empirical experiment. We are making many assumptions favouring T1, but when the result comes in it supports T2 in astonishingly non-trivial ways. The powerset approach provides support for PME against T1 because it combines the assumptions grounding T1 with a limit construction and still yields a solution that closely approximates the one generated by PME rather than the one generated by T1.

On its own the powerset approach is just what Grünwald and Halpern call a naive space, for which CAR does not hold. Hence the powerset approach will not give us a precise solution for the problem, although it may with some plausibility guide us in the right direction—especially if despite all its independence and uniformity assumptions it significantly disagrees with intuition T1.

Let us assume a partition of the Blue and Red territories into sets of equal measure (this is the division into rectangles of scenario III). (MAP) dictates that the number of sets covering A_3 equals the number of sets covering $A_1 \cup A_2$. Initially, any subset of this partition is a candidate for Judy Benjamin to consider. The constraint imposed by (HDQ) is that now we only consider subsets for which there are three times as many partition sets (or rectangles, although we are not necessarily limiting ourselves to rectangles) in A_2 as there are in A_1. In figures 7.5 and 7.6 there are diagrams of two subsets. One of them (figure 7.5) is not a candidate, the other one (figure 7.6)

is.

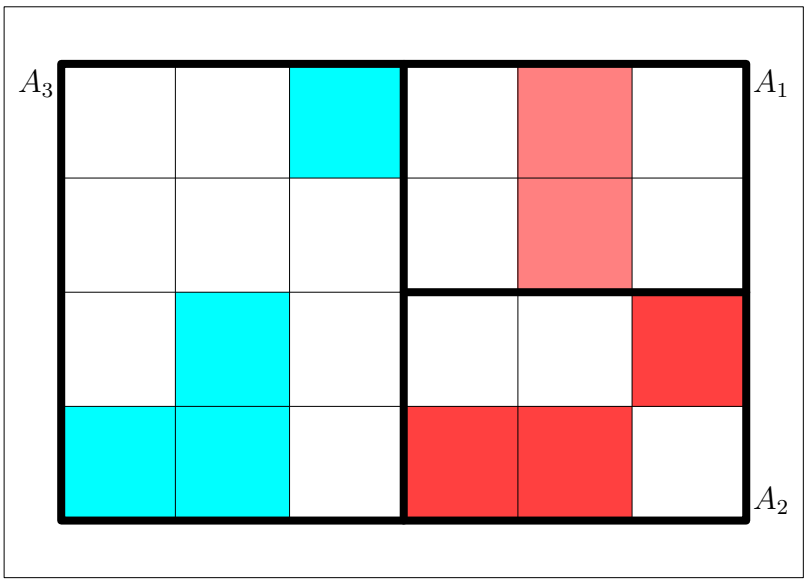

Figure 7.5: This choice of rectangles is not a candidate because the number of rectangles in A_2 is not a t-multiple of the number of rectangles in A_1, here with $s = 2, t = 3$ as in scenario III.

Let X be the random variable that corresponds to the ratio of the number of partition elements (rectangles) that are in A_3 to the total number of partition elements (rectangles) for a randomly chosen candidate. We would now anticipate that the expectation of X (which we will call EX) gives us an indication of the updated probability that Judy is in A_3 (so $EX \approx q_3$). The powerset approach is often superior to the uniformity approach (Grove and Halpern use uniformity, with all the necessary qualifications): if you have played Monopoly, you will know that the frequencies for rolling a 2, a 7, or a 10 with two dice tend to conform more closely to the binomial distribution (based on a powerset approach) rather than to the uniform distribution with $P(\text{rolling } i) = 1/11$ for $i = 2, \ldots, 12$.

Appendix E provides a formula for the powerset approach corresponding to the formula for the PME approach, giving us q_3 dependent on t. Notice that this formula

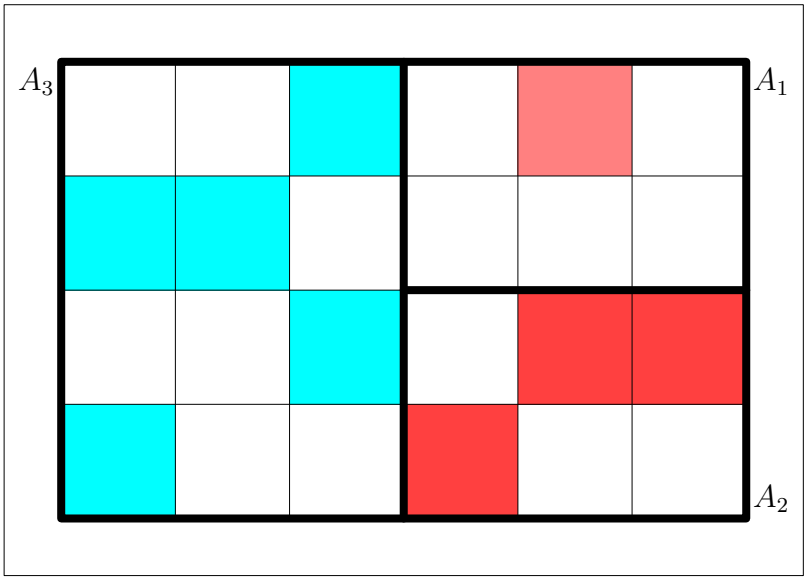

Figure 7.6: This choice of rectangles is a candidate because the number of rectangles in A_2 is a t-multiple of the number of rectangles in A_1, here with $s = 2, t = 3$ as in scenario III.

is for $t = 2, 3, 4, \ldots$. For $t = 1$ use the Chu-Vandermonde identity to find that

$$EX_{12} = (t+1)\frac{\sum_{i=1}^{s} i\binom{ts}{i}\binom{ts}{ti}}{\sum_{i=0}^{s}\binom{ts}{i}\binom{ts}{ti}} = (t+1)\frac{s}{2} \tag{7.10}$$

and consequently $EX = 1/2$, as one would expect. For $t = 1/2, 1/3, 1/4, \ldots$ we can simply reverse the roles of A_1 and A_2. These results give us G_{pws} and a graph of the normalized odds vector (see figure 7.7), a bit bumpy around the middle because the t-values are discrete and farther apart in the middle, as $t = \vartheta/(1-\vartheta)$. Comparing the graphs of the normalized odds vector under Grove and Halpern's uniformity approach (G_{ind}), Jaynes' PME approach (G_{max}), and the powerset approach suggested in this chapter (G_{pws}), it is clear that the powerset approach supports PME.

Going through the calculations, it seems at many places that the powerset approach should give its support to Grove and Halpern's uniformity approach in keeping with intuition T1. It is unexpected to find out that in the mathematical analysis the parameters converge to a non-trivial factor and do not tend to negative or positive

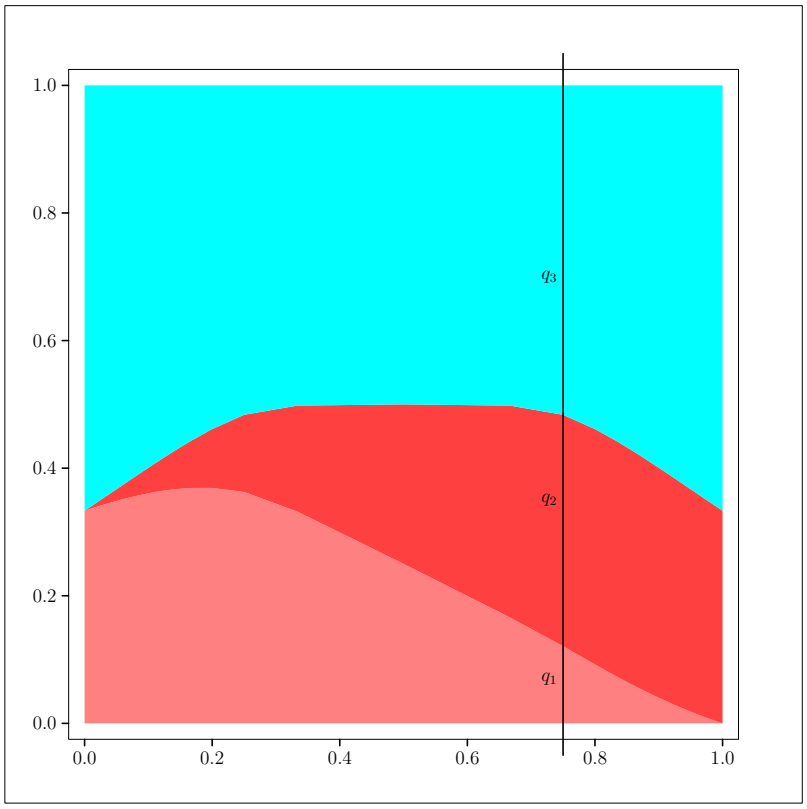

Figure 7.7: Judy Benjamin's updated probability assignment according to the powerset approach. $0 < \vartheta < 1$ forms the horizontal axis, the vertical axis shows the updated probability distribution (or the normalized odds vector) (q_1, q_2, q_3). The vertical line at $\vartheta = 0.75$ shows the specific updated probability distribution G_{pws} for the Judy Benjamin problem.

infinity. Most surprisingly, the powerset approach, prima facie unrelated to an approach using information, supports the idea that a set of events about which nothing is known (such as A_3) gains in probability in the updated probability distribution compared to the set of events about which something is known (such as A_1 and A_2), even if it is only partial information. Unless independence is specified, as in example 39, the area of ignorance gains compared to the area of knowledge.

We now have several ways to characterize Judy's updated probabilities and updated probabilities following upon partial information in general. Only one of them,

the uniformity approach, violates van Fraassen, Hughes, and Harman's five symmetry requirements in (1986) and intuition T2. The uniformity approach, however, is the only one that satisfies intuition T1, an intuition which most people have when they first hear the story.

Two arguments attenuate the position of the uniformity approach in comparison with the others. First, T1 rests on an independence assumption which is not reflected in the problem. Although there is no indication that what Judy's commanders tell her is in any way dependent on her probability of being in Blue territory, it is not excluded either (see scenarios II and III earlier in this chapter). PME takes this uncertainty into consideration. Second, when we investigate the problem using the powerset approach it turns out that a division into equally probable, independent, and increasingly fine bits of information supports not intuition T1 but rather intuition T2. PME, for now, is vindicated. We need to look for full employment not by cleverly manipulating prior probabilities, but by making fresh observations, designing better experiments, and partitioning the theory space more finely.

Chapter 8

Conclusion

It is one goal of formal epistemology to provide an account of normativity for partial beliefs. Bayesian epistemology, for example, defends the view that a rational agent updates relatively prior probabilities to posterior probabilities using standard conditioning if the event on which the agent conditions has a non-zero prior probability. If this probability is zero we get the Borel-Kolmogorov paradox (for a thorough treatment see the soon-to-be-published article "Conditioning using Conditional Expectations: The Borel-Kolmogorov Paradox" by Zalán Gyenis, Miklós Rédei, and Gábor Hofer-Szabó). In this dissertation, I have exclusively addressed finite event spaces so that the Borel-Kolmogorov paradox is not an issue. It is an open question whether Gyenis, Redei, and Hofer-Szabó's treatment provides us with a compact Lebesgue space that is a natural extension of the geometry of reason. In personal communication, Redei speculated that this may be the case. Extending PME to cases where the cardinality of the event space is no longer finite would need to face these questions.

Bayesian epistemology is controversial even where the event space is finite and the probability of the event on which the posterior probability is conditioned is non-zero. It is also the most substantial normative account available for updating probabilities. Accounts that either underdetermine it (by claiming that it requires too much) or overdetermine it (by claiming that it requires too little) are generally in the shadow of the Bayesian account: they either take a few things or add a few things, but they seldom claim the kind of overarching explanatory scope and formal unity inherent in the Bayesian method.

Often this is interpreted to be a good thing: epistemology, after all, is a messy affair, and so it is not surprising that the account with the most formal unity is not

the account that best accords with our intuitions, for that would just be a strange coincidence. There is now a compendium of rules and methods, some of which go beyond Bayesian epistemology while others question its general validity. At the same time, a few formal accounts have emerged which rival Bayesian epistemology in substance: AGM theory and ranking theory, which address full beliefs rather than partial beliefs but have the potential of informing partial belief epistemology as well; Dempster-Shafer theory and possibility theory, which seek to address situations where Bayesian epistemology seems to give the wrong answers.

This dissertation defends a formal account which also rivals Bayesian epistemology in scope and unity, but incorporates it by generalizing its requirements. Whereas Dempster-Shafer theory and possibility theory are alternatives to Bayesian epistemology, information theory agrees with the Bayesian method on all questions that come within the purview of both. What distinguishes information theory from Bayesian epistemology is that information theory is normative where Bayesian epistemology is silent and that information theory has a story of why the Bayesian method works so well. For it turns out that standard conditioning just is the conditioning that minimizes the cross-entropy between the relatively prior probability distribution and the posterior probability distribution. Information theory can now go further: Jeffrey conditioning also minimizes this cross-entropy, and so does, quite obviously, the constraint rule for cross-entropy in cases where standard conditioning and Jeffrey conditioning do not apply.

After the introduction in chapter 1 and an explanation of information theory and entropy in chapter 2, I introduce the partial belief model and its formal properties using information theory in chapter 3. After chapter 3, there is a series of conflicts between information theory and current epistemological accounts. The first such conflict is between information theory and the geometry of reason in chapter 4. Many formal epistemologists now seek ways in which normativity about partial beliefs can be grounded in epistemic virtue. In science, we naturally follow the maxim to be as close as possible to the truth. For partial beliefs, however, the grounding virtues were often thought to be pragmatic or psychological in nature. James Joyce's norm of gradational accuracy has provided a formal tool to make partial beliefs about

epistemic virtue.

Unfortunately, the project has also relied heavily on a Euclidean metric for the difference between partial beliefs. Leitgeb and Pettigrew have correctly pointed out how this must lead to a rejection of Jeffrey conditioning as a natural extension of standard conditioning. Chapter 4 makes the case that given the choice between Jeffrey conditioning and the Euclidean metric, it is the Euclidean metric which should have to go. The story ends with a bitter note and with a sweet note, but most of all with a formidable research project: The bitter note is that not only the geometry of reason violates several strong epistemic intuitions, but so does information theory. The sweet note is that there is hope that Joyce's norm of gradational accuracy will do just as well if it is not built on a Euclidean metric. I am hopeful that both of these open questions can be resolved, and I suspect that the theory of differential manifolds has the tools to resolve them for us.

The next conflict is in chapter 5 between a natural extension of Jeffrey conditioning, which I have called Wagner conditioning, and information theory. According to Wagner, this natural extension is inconsistent with information theory. I have found that his claims are correct, but only if we accept that rational agents sometimes have indeterminate credal states. Considering that the rationality of indeterminate credal states are now well-established in the Bayesian community, this would be a substantial problem for information theory. In chapter 6, I make the case that given the choice to reject information theory or indeterminate credal states, we should reject indeterminate credal states. The reasons are independent of information theory (one could also reject both information theory and indeterminate credal states). Indeterminate credal states, as attractive as they look at first glance, exact a heavy price for the solutions they provide to address the problems of sharp credences.

Finally, in chapter 7, I address a counterexample to information theory that is often invoked by opponents as undermining the case for information theory as a general problem solver in partial belief kinematics. The *Judy Benjamin* case is instructive in different ways, clarifying also the relationship between information theory and epistemic entrenchment, a concept originally tied to AGM belief revision but now increasingly also applied to partial beliefs; and the relationship between information

theory and independence. Information theory appears to favour independence, often because independence is the middle ground between dependence relationships leaning one way or another. Having a bias with respect to dependence would provide information that is often not warranted. Information theory therefore settles on independence, without however informing that a relationship is indeed independent. This ambiguity between lack of commitment to one or another way of dependence and independence is something that information theory must, and does, navigate carefully.

Now that the work is done, it remains to spell out where the open questions and the future research projects are. I already made reference to two of them: (1) It is imperative for information theory to give an account of the violations in List Two in subsection 4.2.3. (2) It is imperative for information theory to give a positive account of what follows from Joyce's norm of gradational accuracy once we strip it of its Euclidean prejudice.

To these two projects I add the following: (3) The precision-modesty paradox briefly mentioned in section 1.2 has significant potential to address a vexing question: the more modest and nuanced we are in our beliefs, the more information we appear to have about the propositions in question. I am interested in what partial beliefs are vis-à-vis full beliefs. Are partial beliefs just a particular kind of full belief? Do full beliefs and partial beliefs coordinate to produce the doxastic state of an agent? Or should we separate partial beliefs from belief psychology and interpret them as representative of an agent's preferences? An answer to this question may inform a solution to the precision-modesty paradox, and vice versa.

(4) While I was working on indeterminate credal states, I found out that an agent using lower and upper previsions (provided by indeterminate credal states) may systematically do better accepting and rejecting bets than an agent who uses sharp credences. Walley conducted an experiment in which this appeared to be true, betting on soccer games played in the Soccer World Cup 1982 in Spain (see Walley, 1991, Appendix I). I replicated the experiment using two computer players with rudimentary artificial intelligence and made them specify betting parameters (previsions) for games played in the Soccer World Cup 2014 in Brazil. I used the Poisson distribution

(which is an excellent predictor for the outcome of soccer matches) and the FIFA ranking to simulate millions of counterfactual World Cup results and their associated bets, using Walley's evaluation method. The player using upper and lower previsions had a slight but systematic advantage over the player using sharp credences. I would love to see the preliminary data explained by analytic expressions of the respective gains by the two players and a successful defence of sharp credences given this at first glance embarrassing state of affairs. I have an idea of how this could be done. It would also be interesting to see how an explanatory account of these dynamics corresponds to the literature in economics on prediction markets and stock market trading.

(5) There are two impossibility theorems which would be of great interest if they could be established. The first one concerns desiderata S1-7 in section 6.2 for an information measure that has as its domain indeterminate rather than sharp credal states. Currently there is no known measure which fulfills these desiderata, but there is also no proof that it does not exist. If it did exist, it could undermine my claim that in terms of conservation of information, indeterminate credal states do not outperform sharp credences, even though this claim is often made. The other impossibility theorem concerns difference measures of probability distributions. I have listed some desiderata in List One and List Two in subsection 4.2.3. It is currently not clear if there is a difference measure that fulfills all of these desiderata, or if there is a proof that no such measure exists. I would also be interested in a measure which supports Jeffrey conditioning while not committing all the violations of information theory in List Two. Leitgeb and Pettigrew suggest that non-Euclidean but symmetric difference measures may have interesting properties, but I have not seen this avenue pursued in the literature.

Finally, a word about what motivated this dissertation on a more philosophical level. I used to feel kinship with some of the idealists in the Brigadas Internacionales of the Spanish Civil War. It was not because I shared any of their lived reality, but because I, as they did, believed in a guiding hand in history and I, as they were, was eventually disappointed in this belief. Karl Popper gives a systematic account of the poverty of historicism in a book carrying that title, based on personal experience

at a street battle in the Hörlstraße in Vienna, where he overheard fellow Marxists endangering the lives of their friends in order to hurry along a revolution which they felt assured would come one way or another. In his *Third Essay* on the *Genealogy of Morals*, Nietzsche puts it nicely:

> Regarding nature as though it were a proof of God's goodness and providence; interpreting history in honour of divine reason, as a constant testimonial to an ethical world order and ethical ultimate purpose; explaining all one's own experiences in the way pious folk have done for long enough, as though everything were providence, a sign, intended, and sent for the salvation of the soul: now all that is over.

I prefer the clockwork mechanism of information theory over the probing genius of epistemologists because the probing genius of epistemologists has at its foundation an optimism about the calibration of our epistemological sensitivities that I do not share. My work in formal epistemology has a larger philosophical point in mind which more recognizably matters than a dissertation on Bayesian updating and maximum entropy: epistemological pessimism, the view that the increasing irrelevance of metaphysical questions in the 20th century is matched by an increasing irrelevance of epistemological questions now. Not only do we not know what there is, we do not know what we know. Humans are living beings, not knowing beings. If knowledge defines them, it does so in the sense that humans are deceived beings. Surrounded by metaphysical and epistemological illusions, humans are by their constitution deluded. Their delusions are what makes them interesting and tragic, creative and humorous.

Skeptics require that we know nothing at all, whereas epistemological pessimists observe that we may know a little, but whatever it is we know we don't know that we know it, and it is not much in comparison to what we could know or what is out there to know for creatures that are better at knowing than we are. Epistemological pessimism is the view that recognizes that we do not drive with a good view of our surroundings. We are driving a car in the fog, at full speed, with no access to the brakes and little access to information about what is going on, while constantly having to make life-changing decisions.

So I end with Laplace's last words, according to Joseph Fourier's account pronounced with difficulty: "Ce que nous connaissons est peu de chose, ce que nous ignorons est immense" (Joseph Fourier, Éloge historique de M. le Marquis de Laplace, delivered 15 June 1829).

Bibliography

Alchourrón, Carlos, Peter Gärdenfors, and David Makinson. "On the Logic of Theory Change: Partial Meet Contraction and Revision Functions." *The Journal of Symbolic Logic* 50, 2: (1985) 510–530.

Amari, Shun-ichi, and Hiroshi Nagaoka. *Methods of Information Geometry*. Providence, RI: American Mathematical Society, 2000.

Anton, Howard, and Robert Busby. *Contemporary Linear Algebra*. New York, NY: Wiley, 2003.

Armendt, Brad. "Is There a Dutch Book Argument for Probability Kinematics?" *Philosophy of Science* 47, 4: (1980) 583–588.

Atkinson, David. "Confirmation and Justification." *Synthese* 184, 1: (2012) 49–61.

Augustin, Thomas. "On the Suboptimality of the Generalized Bayes Rule and Robust Bayesian Procedures from the Decision Theoretic Point of View—a Cautionary Note on Updating Imprecise Priors." In *ISIPTA*. 2003, volume 3, 31–45.

Bar-Hillel, Yehoshua, and Rudolf Carnap. "Semantic information." *The British Journal for the Philosophy of Science* 4, 14: (1953) 147–157.

Belot, Gordon. "Bayesian Orgulity." *Philosophy of Science* 80, 4: (2013) 483–503.

Bertsekas, Dimitri. *Constrained Optimization and Lagrange Multiplier Methods*. Boston, MA: Academic, 1982.

Bolker, Ethan. "Functions Resembling Quotients of Measures." *Transactions of the American Mathematical Society* 124, 2: (1966) 292–312.

———. "A Simultaneous Axiomatization of Utility and Subjective Probability." *Philosophy of Science* 34, 4: (1967) 333–340.

Boltzmann, Ludwig. *On the Nature of Gas Molecules*. Taylor and Francis, 1877.

Boole, George. *An Investigation of the Laws of Thought: On Which Are Founded the Mathematical Theories of Logic and Probabilities*. Walton and Maberly, 1854.

Bovens, Luc. "Judy Benjamin is a Sleeping Beauty." *Analysis* 70, 1: (2010) 23–26.

Bradley, Richard. "Radical Probabilism and Bayesian Conditioning." *Philosophy of Science* 72, 2: (2005) 342–364.

Bradley, Seamus, and Katie Steele. "Uncertainty, Learning, and the Problem of Dilation." *Erkenntnis* 79, 6: (2014) 1287–1303.

———. "Can Free Evidence Be Bad? Value of Information for the Imprecise Probabilist." *Philosophy of Science* 83, 1: (2016) 1–28.

Buck, B., and V.A. Macaulay. *Maximum Entropy in Action: A Collection of Expository Essays*. New York: Clarendon, 1991.

Calin, Ovidiu, and Constantin Udriste. *Geometric Modeling in Probability and Statistics*. Heidelberg, Germany: Springer, 2014.

van Campenhout, Jan, and Thomas Cover. "Maximum Entropy and Conditional Probability." *Information Theory, IEEE Transactions on* 27, 4: (1981) 483–489.

Carnap, Rudolf. *The Continuum of Inductive Methods*. University of Chicago, 1952.

———. *Logical Foundations of Probability*. University of Chicago, 1962.

Carnap, Rudolf, and Yehoshua Bar-Hillel. *An Outline of a Theory of Semantic Information*. Cambridge, MA: MIT, 1952.

Caticha, Ariel. "Entropic Inference and the Foundations of Physics." *Brazilian Chapter of the International Society for Bayesian Analysis* .

Caticha, Ariel, and Adom Giffin. "Updating Probabilities." In *MaxEnt 2006, the 26th International Workshop on Bayesian Inference and Maximum Entropy Methods*. 2006.

Chandler, Jake. "Subjective Probabilities Need Not Be Sharp." *Erkenntnis* 79, 6: (2014) 1273–1286.

Chentsov, Nikolai. *Statistical Decision Rules and Optimal Inference*. Providence, R.I: American Mathematical Society, 1982.

Chino, Naohito. "A Graphical Technique for Representing the Asymmetric Relationships Between N Objects." *Behaviormetrika* 5, 5: (1978) 23–44.

———. "A Generalized Inner Product Model for the Analysis of Asymmetry." *Behaviormetrika* 17, 27: (1990) 25–46.

Chino, Naohito, and Kenichi Shiraiwa. "Geometrical Structures of Some Non-Distance Models for Asymmetric MDS." *Behaviormetrika* 20, 1: (1993) 35–47.

Christensen, David. "Measuring Confirmation." *The Journal of Philosophy* 96, 9: (1999) 437–461.

Coombs, Clyde H. *A Theory of Data*. New York, NY: Wiley, 1964.

Cover, T.M., and J.A. Thomas. *Elements of Information Theory*, volume 6. Hoboken, NJ: Wiley, 2006.

Cox, Richard. "Probability, Frequency and Reasonable Expectation." *American Journal of Physics* 14: (1946) 1.

Coxon, Anthony. *The User's Guide to Multidimensional Scaling*. Exeter, NH: Heinemann Educational Books, 1982.

Crupi, Vincenzo, and Katya Tentori. "State of the Field: Measuring Information and Confirmation." *Studies in History and Philosophy of Science Part A* 47: (2014) 81–90.

Crupi, Vincenzo, Katya Tentori, and Michel Gonzalez. "On Bayesian Measures of Evidential Support: Theoretical and Empirical Issues." *Philosophy of Science* 74, 2: (2007) 229–252.

Csiszár, Imre. "Information-Type Measures of Difference of Probability Distributions and Indirect Observations." *Studia Scientiarum Mathematicarum Hungarica* 2: (1967) 299–318.

Csiszár, Imre, and Paul C Shields. *Information Theory and Statistics: A Tutorial*. Hanover, MA: Now Publishers, 2004.

Cyranski, John F. "Measurement, Theory, and Information." *Information and Control* 41, 3: (1979) 275–304.

De Finetti, B. "Sul Significato Soggettivo della Probabilità." *Fundamenta mathematicae* 17, 298-329: (1931) 197.

De Finetti, Bruno. "La prévision: ses lois logiques, ses sources subjectives." In *Annales de l'institut Henri Poincaré*. Presses universitaires de France, 1937, volume 7, 1–68.

Debbah, Mérouane, and Ralf Müller. "MIMO Channel Modeling and the Principle of Maximum Entropy." *IEEE Transactions on Information Theory* 51, 5: (2005) 1667–1690.

Dempster, Arthur. "Upper and Lower Probabilities Induced by a Multivalued Mapping." *The Annals of Mathematical Statistics* 38, 2: (1967) 325–339.

Diaconis, Persi, and Sandy Zabell. "Updating Subjective Probability." *Journal of the American Statistical Association* 77, 380: (1982) 822–830.

Dias, P., and A. Shimony. "A Critique of Jaynes' Maximum Entropy Principle." *Advances in Applied Mathematics* 2: (1981) 172–211.

Domotor, Zoltan. "Probability Kinematics and Representation of Belief Change." *Philosophy of Science* 47, 3: (1980) 384–403.

Döring, Frank. "Why Bayesian Psychology Is Incomplete." *Philosophy of Science* 66: (1999) S379–S389.

Douven, I., and J.W. Romeijn. "A New Resolution of the Judy Benjamin Problem." *CPNSS Working Paper* 5, 7: (2009) 1–22.

Douven, Igor, and Timothy Williamson. "Generalizing the Lottery Paradox." *British Journal for the Philosophy of Science* 57, 4: (2006) 755–779.

Earman, John. *Bayes or Bust?* Cambridge, MA: MIT, 1992.

Edwards, Ward, Harold Lindman, and Leonard J. Savage. "Bayesian Statistical Inference for Psychological Research." *Psychological Review* 70, 3: (1963) 193.

Elga, Adam. "Subjective Probabilities Should Be Sharp." *Philosophers' Imprint* 10, 5: (2010) 1–11.

Ellsberg, Daniel. "Risk, Ambiguity, and the Savage Axioms." *The Quarterly Journal of Economics* 75, 4: (1961) 643–669.

Kampé de Fériet, J., and B. Forte. "Information et probabilité." *Comptes rendus de l'Académie des sciences* A 265: (1967) 110–114.

Festa, R. "Bayesian Confirmation." In *Experience, Reality, and Scientific Explanation: Essays in Honor of Merrilee and Wesley Salmon*, edited by Merrilee H. Salmon, Maria Carla Galavotti, and Alessandro Pagnini, Dordrecht: Kluwer, 1999, 55–88.

Fitelson, Branden. "A Bayesian Account of Independent Evidence with Applications." *Philosophy of Science* 68, 3: (2001) 123–140.

———. "Logical Foundations of Evidential Support." *Philosophy of Science* 73, 5: (2006) 500–512.

Foley, Richard. *Working Without a Net: A Study of Egocentric Epistemology.* New York, NY: Oxford University Press, 1993.

Forster, Malcolm, and Elliott Sober. "How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions." *The British Journal for the Philosophy of Science* 45, 1: (1994) 1–35.

van Fraassen, Bas. "A Problem for Relative Information Minimizers in Probability Kinematics." *The British Journal for the Philosophy of Science* 32, 4: (1981) 375–379.

———. "Belief and the Will." *Journal of Philosophy* 81, 5: (1984) 235–256.

———. *Laws and Symmetry*. Oxford, UK: Clarendon, 1989.

———. "Symmetries of Probability Kinematics." In *Bridging the Gap: Philosophy, Mathematics, and Physics*, Springer, 1993, 285–314.

van Fraassen, Bas, R.I.G. Hughes, and Gilbert Harman. "A Problem for Relative Information Minimizers, Continued." *The British Journal for the Philosophy of Science* 37, 4: (1986) 453–463.

Friedman, Kenneth, and Abner Shimony. "Jaynes's Maximum Entropy Prescription and Probability Theory." *Journal of Statistical Physics* 3, 4: (1971) 381–384.

Gage, Douglas W, and David Hestenes. "Comment on the Paper Jaynes's Maximum Entropy Prescription and Probability Theory." *Journal of Statistical Physics* 7, 1: (1973) 89–90.

Gaifman, Haim. "Subjective Probability, Natural Predicates and Hempel's Ravens." *Erkenntnis* 14, 2: (1979) 105–147.

Gemes, Ken. "Verisimilitude and Content." *Synthese* 154, 2: (2007) 293–306.

Gentleman, R., B. Ding, S. Dudoit, and J. Ibrahim. "Distance Measures in DNA Microarray Data Analysis." In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, edited by R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit, Springer, 2006.

Gibbs, Josiah Willard. *Elementary Principles in Statistical Physics*. New Haven, CT: Yale University, 1902.

Giffin, Adom. *Maximum Entropy: The Universal Method for Inference*. PhD dissertation, University at Albany, State University of New York, Department of Physics, 2008.

Gillies, Donald. *Philosophical Theories of Probability*. London, UK: Routledge, 2000.

Goldstein, Michael. "The Prevision of a Prevision." *Journal of the American Statistical Association* 78, 384: (1983) 817–819.

Good, Irving. *Good Thinking: The Foundations of Probability and Its Applications*. Minneapolis, MN: University of Minnesota, 1983.

Good, John Irving. *Probability and the Weighing of Evidence*. London, UK: Griffin, 1950.

Greaves, Hilary, and David Wallace. "Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility." *Mind* 115, 459: (2006) 607–632.

Grove, A., and J.Y. Halpern. "Probability Update: Conditioning Vs. Cross-Entropy." In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*. 1997, 208–214.

Grünwald, Peter. "Maximum Entropy and the Glasses You Are Looking Through." In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann, 2000, 238–246.

Grünwald, Peter, and Joseph Halpern. "Updating Probabilities." *Journal of Artificial Intelligence Research* 19: (2003) 243–278.

Guiaşu, Silviu. *Information Theory with Application*. New York, NY: McGraw-Hill, 1977.

Gyenis, Zalán, and Miklós Rédei. "Defusing Bertrand's Paradox." *The British Journal for the Philosophy of Science* 66, 2: (2015) 349–373.

Hacking, Ian. "Slightly More Realistic Personal Probability." *Philosophy of Science* 34, 4: (1967) 311–325.

Hájek, Alan. "What Conditional Probability Could Not Be." *Synthese* 137, 3: (2003) 273–323.

Hájek, Alan, and James Joyce. "Confirmation." In *Routledge Companion to the Philosophy of Science*, edited by S. Psillos, and M. Curd, New York, NY: Routledge, 2008, 115–129.

Hájek, Alan, and Michael Smithson. "Rationality and Indeterminate Probabilities." *Synthese* 187, 1: (2012) 33–48.

Halpern, Joseph. "A Counterexample to Theorems of Cox and Fine." *Journal of Artifcial Intelligence Research* 10: (1999) 67–85.

———. *Reasoning About Uncertainty*. Cambridge, MA: MIT, 2003.

Harshman, Richard, and Margaret Lundy. "The PARAFAC Model for Three-Way Factor Analysis and Multidimensional Scaling." In *Research methods for multimode data analysis*, edited by Henry G. Law, New York, NY: Praeger, 1984, 122–215.

Harshman, Richard A., Paul E. Green, Yoram Wind, and Margaret E. Lundy. "A Model for the Analysis of Asymmetric Data in Marketing Research." *Marketing Science* 1, 2: (1982) 205–242.

Hempel, Carl G., and Paul Oppenheim. "A Definition of Degree of Confirmation." *Philosophy of science* 12, 2: (1945) 98–115.

Hobson, A. *Concepts in Statistical Mechanics*. New York, NY: Gordon and Beach, 1971.

Hobson, Arthur. "The Interpretation of Inductive Probabilities." *Journal of Statistical Physics* 6, 2: (1972) 189–193.

Howson, Colin, and Allan Franklin. "Bayesian Conditionalization and Probability Kinematics." *The British Journal for the Philosophy of Science* 45, 2: (1994) 451–466.

Howson, Colin, and Peter Urbach. *Scientific Reasoning: The Bayesian Approach, Third Edition*. Chicago: Open Court, 2006.

Huisman, Leendert. "On Indeterminate Updating of Credences." *Philosophy of Science* 81, 4: (2014) 537–557.

Huttegger, Simon. "Merging of Opinions and Probability Kinematics." *The Review of Symbolic Logic* 8: (2015) 611–648.

Ingarden, R. S., and K. Urbanik. "Information Without Probability." *Colloquium Mathematicum* 9: (1962) 131–150.

Jaynes, E.T. "Information Theory and Statistical Mechanics." *Physical Review* 106, 4: (1957a) 620–630.

———. "Information Theory and Statistical Mechanics II." *Physical Review* 108, 2: (1957b) 171.

———. "The Well-Posed Problem." *Foundations of Physics* 3, 4: (1973) 477–492.

———. "Where Do We Stand on Maximum Entropy." In *The Maximum Entropy Formalism*, edited by R.D. Levine, and M. Tribus, Cambridge, MA: MIT, 1978, 15–118.

———. "Some Random Observations." *Synthese* 63, 1: (1985) pp. 115–138.

———. "Optimal Information Processing and Bayes's Theorem: Comment." *The American Statistician* 42, 4: (1988) 280–281.

———. *Papers on Probability, Statistics and Statistical Physics*. Dordrecht: Springer, 1989.

Jaynes, E.T., and G.L. Bretthorst. *Probability Theory: the Logic of Science*. Cambridge, UK: Cambridge University, 2003.

Jeffrey, Richard. *The Logic of Decision*. New York, NY: McGraw-Hill, 1965.

———. "Dracula Meets Wolfman: Acceptance Versus Partial Belief." In *Induction, Acceptance and Rational Belief*, edited by Marshall Swain, Springer, 1970, 157–185.

———. "Axiomatizing the Logic of Decision." In *Foundations and Applications of Decision Theory*, Springer, 1978, 227–231.

———. "Bayesianism with a Human Face." *Minnesota Studies in the Philosophy of Science* 10: (1983) 133–156.

Jeffreys, Harold. *Scientific Inference*. Cambridge University, 1931.

———. *The Theory of Probability*. Cambridge University, 1939.

Joyce, James. "A Nonpragmatic Vindication of Probabilism." *Philosophy of Science* 65, 4: (1998) 575–603.

———. *The Foundations of Causal Decision Theory*. Cambridge University, 1999.

———. "How Probabilities Reflect Evidence." *Philosophical Perspectives* 19, 1: (2005) 153–178.

———. "The Development of Subjective Bayesianism." In *Handbook of the History of Logic 10*, edited by D. M. Gabbay, S. Hartmann, and J. Woods, Amsterdam, Netherlands: Elsevier, 2009, 415–476.

———. "A Defense of Imprecise Credences in Inference and Decision Making." *Philosophical Perspectives* 24, 1: (2010) 281–323.

Kaplan, Mark. "In Defense of Modest Probabilism." *Synthese* 176, 1: (2010) 41–55.

Karmeshu, J. *Entropy Measures, Maximum Entropy Principle and Emerging Applications*. New York: Springer-Verlag, 2003.

Kelly, T. "The Epistemic Significance of Disagreement." In *Oxford Studies in Epistemology 1*, edited by Tamar Gendler, and John Hawthorne, New York, NY: Oxford University, 2008, 167–196.

Keynes, John Maynard. *A Treatise on Probability*. Diamond, 1909.

———. *A Treatise on Probability*. London, UK: Macmillan, 1921.

Khinchin, A.I. *Mathematical Foundations of Information Theory*. New York, NY: Dover, 1957.

Klir, George. *Uncertainty and Information: Foundations of Generalized Information Theory*. Hoboken, NJ: Wiley, 2006.

Kolmogorov, A.N. "Logical Basis for Information Theory and Probability Theory." *IEEE Transactions on Information Theory* 14, 5: (1968) 662–664.

Kopperman, Ralph. "All Topologies Come from Generalized Metrics." *American Mathematical Monthly* 95, 2: (1988) 89–97.

Kullback, Solomon. *Information Theory and Statistics*. Dover Publications, 1959.

Kullback, Solomon, and Richard Leibler. "On Information and Sufficiency." *The Annals of Mathematical Statistics* 22, 1: (1951) 79–86.

Kyburg, Henry. "Recent Work in Inductive Logic." In *Recent Work in Philosophy*, edited by Tibor Machan, and Kenneth Lucey, Totowa, NJ: Rowman and Allanheld, 1983, 87–150.

———. "Book Review: Betting on Theories by Patrick Maher." *Philosophy of Science* 62, 2: (1995) 343.

Landes, Jürgen, and Jon Williamson. "Objective Bayesianism and the Maximum Entropy Principle." *Entropy* 15, 9: (2013) 3528–3591.

Lange, Marc. "Is Jeffrey Conditionalization Defective by Virtue of Being Non-Commutative?" *Synthese* 123, 3: (2000) 393–403.

Leitgeb, Hannes, and Richard Pettigrew. "An Objective Justification of Bayesianism I: Measuring Inaccuracy." *Philosophy of Science* 77, 2: (2010a) 201–235.

———. "An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy." *Philosophy of Science* 77, 2: (2010b) 236–272.

Levi, Isaac. "Probability Kinematics." *The British Journal for the Philosophy of Science* 18, 3: (1967) 197–209.

———. "Direct Inference and Confirmational Conditionalization." *Philosophy of Science* 48, 4: (1981) 532–552.

———. "Imprecision and Indeterminacy in Probability Judgment." *Philosophy of Science* 52, 3: (1985) 390–409.

———. "The Demons of Decision." *The Monist* 70, 2: (1987) 193–211.

Levinstein, Benjamin Anders. "Leitgeb and Pettigrew on Accuracy and Updating." *Philosophy of Science* 79, 3: (2012) 413–424.

Lewis, David. "A Subjectivist's Guide to Objective Chance." In *Ifs: Conditionals, Belief, Decision, Chance and Time*, edited by William Harper, Robert Stalnaker, and G.A. Pearce, Springer, 1981, 267–297.

———. "Many, But Almost One." In *Ontology, Causality, and Mind: Essays on the Philosophy of D.M. Armstrong*, edited by Keith Campbell, Bacon Bacon, and Lloyd Reinhardt, Cambridge, UK: Cambridge University, 1993, 23–38.

———. "Why Conditionalize." In *Papers in Metaphysics and Epistemology: Volume 2*, Cambridge University, 2010, 403–407.

Lukits, Stefan. "The Principle of Maximum Entropy and a Problem in Probability Kinematics." *Synthese* 191, 7: (2014b) 1409–1431.

———. "Maximum Entropy and Probability Kinematics Constrained by Conditionals." *Entropy* Special Issue "Maximum Entropy Applied to Inductive Logic and Reasoning": (2015) forthcoming.

MacKay, David. *Information Theory, Inference and Learning Algorithms*. Cambridge, UK: Cambridge, 2003.

Maher, Patrick. "Diachronic Rationality." *Philosophy of Science* 120–141.

———. *Betting on Theories*. Cambridge University, 1993.

Majerník, Vladimír. "Marginal Probability Distribution Determined by the Maximum Entropy Method." *Reports on Mathematical Physics* 45, 2: (2000) 171–181.

Mikkelson, Jeffrey. "Dissolving the Wine/Water Paradox." *The British Journal for the Philosophy of Science* 55, 1: (2004) 137–145.

Miller, David. "A Geometry of Logic." In *Aspects of Vagueness*, edited by Heinz Skala, Settimo Termini, and Enric Trillas, Dordrecht, Holland: Reidel, 1984, 91–104.

Milne, Peter. "log[P(h/eb)/P(h/b)] Is the One True Measure of Confirmation." *Philosophy of Science* 63, 1: (1996) 21–26.

———. "Information, Confirmation, and Conditionals." *Journal of Applied Logic* 12, 3: (2014) 252–262.

von Mises, Richard. *Mathematical Theory of Probability and Statistics*. New York, NY: Academic, 1964.

Mork, Jonas Clausen. "Uncertainty, Credal Sets and Second Order Probability." *Synthese* 190, 3: (2013) 353–378.

Mormann, Thomas. "Geometry of Logic and Truth Approximation." *Poznan Studies in the Philosophy of the Sciences and the Humanities* 83, 1: (2005) 431–454.

Moss, Sarah. "Epistemology Formalized." *Philosophical Review* 122, 1: (2013) 1–43.

Neapolitan, Richard E, and Xia Jiang. "A Note of Caution on Maximizing Entropy." *Entropy* 16, 7: (2014) 4004–4014.

Nozick, Robert. *Philosophical Explanations*. Cambridge, MA: Harvard University, 1981.

Oddie, Graham. "The Content, Consequence and Likeness Approaches to Verisimilitude: Compatibility, Trivialization, and Underdetermination." *Synthese* 190, 9: (2013) 1647–1687.

van Ommen, Sandrien, Petra Hendriks, Dicky Gilbers, Vincent van Heuven, and Charlotte Gooskens. "Is Diachronic Lenition a Factor in the Asymmetry in Intelligibility Between Danish and Swedish?" *Lingua* 137: (2013) 193–213.

Paris, Jeff. *The Uncertain Reasoner's Companion: A Mathematical Perspective*, volume 39. Cambridge, UK: Cambridge University, 2006.

Paris, Jeff, and Alena Vencovská. "A Note on the Inevitability of Maximum Entropy." *International Journal of Approximate Reasoning* 4, 3: (1990) 183–223.

Pettigrew, Richard. "Epistemic Utility and Norms for Credences." *Philosophy Compass* 8, 10: (2013) 897–908.

Popper, Karl. *Conjectures and Refutations*. London, UK: Routledge and Kegan Paul, 1963.

Ramsey, F.P. "Truth and Probability." In *Foundations of Mathematics and other Essays*, edited by R. B. Braithwaite, London, UK: Kegan, Paul, Trench, Trubner, and Company, 1926, 156–198.

Rinard, Susanna. "A Decision Theory for Imprecise Probabilities." *Philosopher's Imprint* 15, 7: (2015) 1–16.

Rosenkrantz, R.D. "Bayesian Confirmation: Paradise Regained." *British Journal for the Philosophy of Science* 45, 2: (1994) 467–476.

Rowbottom, Darrell. "The Insufficiency of the Dutch Book Argument." *Studia Logica* 87, 1: (2007) 65–71.

Saburi, S., and N. Chino. "A Maximum Likelihood Method for an Asymmetric MDS Model." *Computational Statistics & Data Analysis* 52, 10: (2008) 4673–4684.

Salmon, W.C. *The Foundations of Scientific Inference*. Pittsburgh, PA: University of Pittsburgh, 1967.

Salmon, Wesley. "Confirmation and Relevance." In *Induction, Probability, and Confirmation*, edited by Grover Maxwell, and Robert Milford Anderson, Minneapolis, MI: University of Minnesota Press, 1975, 3–36.

Savage, Leonard. *The Foundations of Statistics*. New York, NY: Wiley, 1954.

Schlesinger, George. "Measuring Degrees of Confirmation." *Analysis* 55, 3: (1995) 208–212.

Seidenfeld, Teddy. "Why I Am Not an Objective Bayesian; Some Reflections Prompted by Rosenkrantz." *Theory and Decision* 11, 4: (1979) 413–440.

———. "Entropy and Uncertainty." In *Advances in the Statistical Sciences: Foundations of Statistical Inference*, Springer, 1986, 259–287.

Seidenfeld, Teddy, Mark Schervish, and Joseph Kadane. "When Fair Betting Odds Are Not Degrees of Belief." In *Proceedings of the Biennial Meeting of the Philosophy of Science Association*. JSTOR, 1990, 517–524.

Seidenfeld, Teddy, and Larry Wasserman. "Dilation for Sets of Probabilities." *The Annals of Statistics* 1139–1154.

Sen, Amartya. "Choice Functions and Revealed Preference." *The Review of Economic Studies* 38, 3: (1971) 307–317.

Shannon, C. E. "A Mathematical Theory of Communication." *The Bell System Technical Journal* 27, 1: (1948) 379–423, 623–656.

Shannon, Claude Elwood. "A Mathematical Theory of Communication." *ACM SIGMOBILE Mobile Computing and Communications Review* 5, 1: (2001) 3–55.

Shimony, Abner. "The Status of the Principle of Maximum Entropy." *Synthese* 63, 1: (1985) 35–53.

———. *The Search for a Naturalistic World View*. Cambridge University, 1993.

Shogenji, Tomoji. "The Degree of Epistemic Justification and the Conjunction Fallacy." *Synthese* 184, 1: (2012) 29–48.

Shore, John, and Rodney Johnson. "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy." *IEEE Transactions on Information Theory* 26, 1: (1980) 26–37.

Shun-ichi, Amari. *Differential-Geometrical Methods in Statistics*. Berlin, Germany: Springer, 1985.

Skilling, John. "The Axioms of Maximum Entropy." In *Maximum-Entropy and Bayesian Methods in Science and Engineering*, edited by G.J. Erickson, and C.R. Smith, Dordrecht, Holland: Springer, 1988, 173–187.

Skyrms, Brian. "Maximum Entropy Inference as a Special Case of Conditionalization." *Synthese* 63, 1: (1985) 55–74.

———. "Dynamic Coherence." In *Advances in the Statistical Sciences: Foundations of Statistical Inference*, Springer, 1986, 233–243.

———. "Coherence." In *Scientific Inquiry in Philosophical Perspective*, edited by N. Rescher, Lanham, MD: University Press of America, 1987a, 225–242.

———. "Updating, Supposing, and Maxent." *Theory and Decision* 22, 3: (1987b) 225–246.

Sober, Elliott. "No Model, No Inference: A Bayesian Primer on the Grue Problem." In *Grue!: The New Riddle of Induction*, edited by Douglas Frank Stalker, Chicago: Open Court, 1994, 225–240.

Spohn, Wolfgang. *The Laws of Belief: Ranking Theory and Its Philosophical Applications*. Oxford University, 2012.

Stephens, Christopher. "When Is It Selectively Advantageous to Have True Beliefs? Sandwiching the Better Safe Than Sorry Argument." *Philosophical Studies* 105, 2: (2001) 161–189.

Talbott, W. J. "Two Principles of Bayesian Epistemology." *Philosophical Studies* 62, 2: (1991) 135–150.

Teller, Paul. "Conditionalization and Observation." *Synthese* 26, 2: (1973) 218–258.

———. "Conditionalization, Observation, and Change of Preference." In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, Dordrecht: Reidel, 1976.

Tikochinsky, Y, NZ Tishby, and RD Levine. "Alternative Approach to Maximum-Entropy Inference." *Physical Review A* 30, 5: (1984) 2638.

Tobler, Waldo. *Spatial Interaction Patterns*. Schloss Laxenburg, Austria: International Institute for Applied Systems Analysis, 1975.

Topey, Brett. "Coin Flips, Credences and the Reflection Principle." *Analysis* 72, 3: (2012) 478–488.

Topsøe, F. "Information-Theoretical Optimization Techniques." *Kybernetika* 15, 1: (1979) 8–27.

Tribus, Myron, and Hector Motroni. "Comments on the Paper Jaynes's Maximum Entropy Prescription and Probability Theory." *Journal of Statistical Physics* 4, 2: (1972) 227–228.

Tversky, Amos. "Features of Similarity." *Psychological Review* 84, 4: (1977) 327–352.

Uffink, Jos. "Can the Maximum Entropy Principle Be Explained as a Consistency Requirement?" *Studies in History and Philosophy of Science* 26, 3: (1995) 223–261.

———. "The Constraint Rule of the Maximum Entropy Principle." *Studies in History and Philosophy of Science* 27, 1: (1996) 47–79.

Wagner, Carl. "Probability Kinematics and Commutativity." *Philosophy of Science* 69, 2: (2002) 266–278.

Wagner, Carl G. "Generalized Probability Kinematics." *Erkenntnis* 36, 2: (1992) 245–257.

Walley, Peter. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall London, 1991.

Weatherson, Brian. "Knowledge, Bets and Interests." In *Knowledge Ascriptions*, edited by Jessica Brown, and Mikkel Gerken, Oxford, UK: Oxford University, 2012, 75–103.

———. "For Bayesians, Rational Modesty Requires Imprecision." *Ergo* 2, 20: (2015) 1–16.

Weisberg, Jonathan. "You've Come a Long Way, Bayesians." *Journal of Philosophical Logic* 1–18.

White, Roger. "Evidential Symmetry and Mushy Credence." In *Oxford Studies in Epistemology 3*, edited by Tamar Gendler, and John Hawthorne, New York, NY: Oxford University, 2010, 161–186.

Wiener, Norbert. "The Ergodic Theorem." *Duke Mathematical Journal* 5, 1: (1939) 1–18.

Williams, P.M. "Bayesian Conditionalisation and the Principle of Minimum Information." *British Journal for the Philosophy of Science* 131–144.

Williamson, Jon. "Objective Bayesianism, Bayesian Conditionalisation and Voluntarism." *Synthese* 178, 1: (2011) 67–85.

Williamson, Timothy. *Vagueness*. New York, NY: Routledge, 1996.

———. *Knowledge and Its Limits*. Oxford, UK: Oxford University, 2000.

Zabell, Sandy L. *Symmetry and its Discontents: Essays on the History of Inductive Probability*. Cambridge, UK: Cambridge University, 2005.

Zalabardo, José. "An Argument for the Likelihood-Ratio Measure of Confirmation." *Analysis* 69, 4: (2009) 630–635.

Zellner, Arnold. "Optimal Information Processing and Bayes's Theorem." *The American Statistician* 42, 4: (1988) 278–280.

Zubarev, D.N., V. Morozov, and G. Roepke. *Non-Equilibrium Statistical Thermodynamics*. Whoknows, 1974.

Appendix A

Weak Convexity and Symmetry in Information Geometry

Using information theory instead of the geometry of reason, Joyce's result still stands, vindicating probabilism on epistemic merits rather than prudential ones: partial beliefs which violate probabilism are dominated by partial beliefs which obey it, no matter what the facts are.

Joyce's axioms, however, will need to be reformulated to accommodate asymmetry. This appendix shows that the axiom Weak Convexity (see section 4.2) still holds in information geometry. Consider three points $Q, R, S \in \mathbb{S}^{n-1}$ (replace $\mathbb{S}^{n-1}$ by the n-dimensional space of non-negative real numbers, if you do not want to assume probabilism) for which

$$D_{\mathrm{KL}}(Q, R) = D_{\mathrm{KL}}(Q, S). \tag{A.1}$$

I will show something slightly stronger than Weak Convexity: Joyce's inequality is not only true for the midpoint between R and S but for all points $\vartheta R + (1 - \vartheta)S$, as long as $0 \leq \vartheta \leq 1$. The inequality aimed for is

$$D_{\mathrm{KL}}(Q, \vartheta R + (1 - \vartheta)S) \leq D_{\mathrm{KL}}(Q, R) = D_{\mathrm{KL}}(Q, S). \tag{A.2}$$

To show that it holds I need the log-sum inequality, which is a result of Jensen's inequality (for a proof of the log-sum inequality see Theorem 2.7.1 in Cover and Thomas, 2006, 31). For non-negative numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$,

$$\sum_{i=1}^{n} a_i \ln \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \ln \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}. \tag{A.3}$$

(A.2) follows from (A.3) via

$$\begin{aligned} &D_{\mathrm{KL}}(Q, R) = \vartheta D_{\mathrm{KL}}(Q, R) + (1 - \vartheta) D_{\mathrm{KL}}(Q, S) = \\ &\sum_{i=1}^{n} \left(\vartheta q_i \ln \frac{\vartheta q_i}{\vartheta r_i} + (1 - \vartheta) q_i \ln \frac{(1 - \vartheta) q_i}{(1 - \vartheta) s_i} \right) \geq \\ &\sum_{i=1}^{n} q_i \ln \frac{q_i}{\vartheta r_i + (1 - \vartheta) s_i} = D_{\mathrm{KL}}(Q, \vartheta R + (1 - \vartheta) S). \end{aligned} \tag{A.4}$$

I owe some thanks to physicist friend Thomas Buchberger for help with this proof. Interested readers can find a more general claim in Csiszár's Lemma 4.1 (see Csiszár and Shields, 2004, 448) which accommodates convexity of the Kullback-Leibler divergence as a special case.

Appendix B

Asymmetry in Two Dimensions

This appendix contains a proof that the threefold partition (4.59) of $\mathbb{S}^1$ is well-behaved, in contrast to the threefold partition of $\mathbb{S}^2$ as illustrated by figure B.1. For the two-dimensional case, i.e. considering $p, q \in \mathbb{S}^1$ with $0 < p, q < 1, p + p' = 1$ and $q + q' = 1$,

$$\begin{array}{rcl} \Delta_q(p) > 0 & \text{for} & |p - p'| > |q - q'| \\ \Delta_q(p) = 0 & \text{for} & |p - p'| = |q - q'| \\ \Delta_q(p) < 0 & \text{for} & |p - p'| < |q - q'| \end{array} \tag{B.1}$$

where $\Delta_q(p) = D_{\mathrm{KL}}(q, p) - D_{\mathrm{KL}}(p, q)$ and $D_{\mathrm{KL}}(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1-p}{1-q}$. Part of information theory's ill behaviour outlined in section 4.5.3 is that in the higher-dimensional case the partition does not follow the simple rule that higher entropy of P compared to Q implies that $\Delta_Q(P) > 0$ (Δ here defined as in (4.58)). In the two-dimensional case, however, this simple rule applies.

That a comparison in entropy $H(p) = -p \log p - (1 - p) \log(1 - p)$ between $H(p)$ and $H(q)$ corresponds to a comparison of $|p - p'|$ and $|q - q'|$ is trivial. The proof for (B.1) is straightforward given the following non-trivial lemma establishing a very tight inequality. Given that $p + p' = 1$ and $q + q' = 1$ and $p, q, p', q' > 0$ it is true that

$$\text{If } \log(p/q) > \log(q'/p') \text{ then } (p + q) \log(p/q) > (p' + q') \log(q'/p') \tag{B.2}$$

Let $x = p/q$ and $y = q'/p'$. We know that $x > y$ since $\log x > \log y$. Now we want to

show $(p+q)\log x > (p'+q')\log y$. Note that $p = xq, q' = p'y, p+q = q(x+1)$, and $p'+q' = p'(y+1)$. Therefore,

$$q = \frac{1-y}{1-xy} \tag{B.3}$$

and

$$p' = \frac{1-x}{1-xy}. \tag{B.4}$$

What we want to show is that $x > y$ implies

$$\frac{1-y}{1-xy}(x+1)\log x > \frac{1-x}{1-xy}\log y. \tag{B.5}$$

Note that $f(x) = (1-x)^{-1}(x+1)\log x$ is increasing on $(0,1)$ and decreasing on $(1,\infty)$, and consider the following two cases:

(i) When $x < 1, y < 1$, (B.5) follows from the fact that f is increasing on $(0,1)$.

(ii) When $x > 1, y > 1$, (B.5) follows from the fact that f is decreasing on $(1,\infty)$.

Mixed cases such as $x > 1, y < 1$ do not occur, as for example $x > 1$ implies $y > 1$.

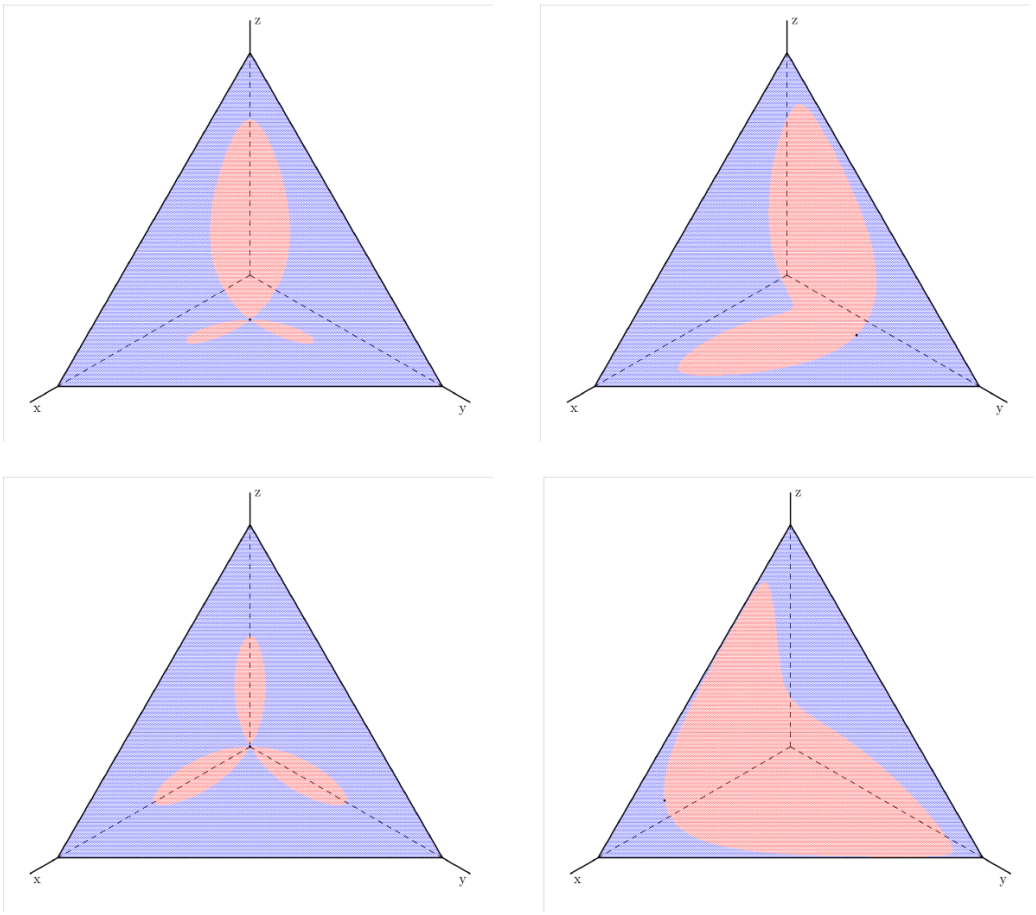

Figure B.1: The partition (4.59) based on different values for P. From top left to bottom right, $P = (0.4, 0.4, 0.2); P = (0.242, 0.604, 0.154); P = (1/3, 1/3, 1/3); P = (0.741, 0.087, 0.172)$. Note that for the geometry of reason, the diagrams are trivial. The challenge for information theory is to explain the non-triviality of these diagrams epistemically without begging the question.

Appendix C

The Horizon Requirement Formalized

Consider the following two conditions on a difference measure D on a simplex $\mathbb{S}^{n-1} \subset \mathbb{R}^n$, which is assumed to be a smooth function from $\mathbb{S}^{n-1} \times \mathbb{S}^{n-1} \to \mathbb{R}$.

(h1) If p, p', q, q' are collinear with the centre of the simplex m (whose coordinates are $m_i = 1/n$ for all i) and an arbitrary but fixed boundary point $\xi \in \partial\mathbb{S}^{n-1}$ and p, p', q, q' are all between m and ξ with $\|p' - p\| = \|q' - q\|$ where p is strictly closest to m, then $|D(p, p')| < |D(q, q')|$. For an illustration of this condition see figure 4.3.

(h2) Let $\mu \in (-1, 1)$ be fixed and D_μ defined as in (C.2). Then $dD_\mu/dx > 0$, where dD_μ/dx is the total derivative as x moves towards $\xi(x)$, the unique boundary point which is collinear with x and m.

To define D_μ, the hardest part is to specify the domain. Let this domain $V(\mu) \subseteq \mathbb{S}^{n-1}$ be defined as

$$V(\mu) = \begin{cases} \{x \in \mathbb{S}^{n-1} | x_i < (1-\mu)\xi_i(x) + \mu m_i, i = 1, \ldots, n\} & \text{for } \mu > 0 \\ \{x \in \mathbb{S}^{n-1} | x_i > (1+\mu)m_i - \mu\xi_i(x), i = 1, \ldots, n\} & \text{for } \mu < 0. \end{cases} \tag{C.1}$$

Then $D_\mu : V(\mu) \to \mathbb{R}_0^+$ is defined as

$$D_\mu(x) = |D(x, y(x))| \tag{C.2}$$

where $y_i(x) = x_i + \mu(\xi_i(x) - m_i)$. Remember that $\xi(x)$ is the unique boundary point which is collinear with x and m. Now for the proof that (h1) and (h2) are equivalent.

First assume (h1) and the negation of (h2). Since D is smooth, there must be a $\bar{\mu}$ and two points x' and x'' collinear with m and a boundary point $\bar{\xi}$ such that $D_{\bar{\mu}}(x') \geq D_{\bar{\mu}}(x'')$ even though $\|\bar{\xi} - x''\| < \|\bar{\xi} - x'\|$. If this were not the case, D_μ would be strictly increasing running towards the boundary points for all μ and its total derivative would be strictly positive so that (h2) follows. Now consider the four points x', x'', y', y'' where $y'_i = x'_i + \mu(\bar{\xi}_i - m_i)$ and $y''_i = x''_i + \mu(\bar{\xi}_i - m_i)$ for $i = 1, \ldots, n$. Without loss of generality, assume $\bar{\mu} > 0$. Then x', x'', y', y'' fulfill the conditions in (h1) and $D_{\bar{\mu}}(x') < D_{\bar{\mu}}(x'')$, in contradiction to the aforesaid.

Then assume (h2). Let x', x'', y', y'' be four points as in (h1). Consider $\mu = \|\xi - m\|/\|x'' - x'\|$. Then $D_\mu(x') = |D(x', x'')|$ and $D_\mu(y') = |D(y', y'')|$. (h2) tells us that along a path from m to ξ, D_μ is strictly increasing, so $D_\mu(x') = |D(x', x'')| < |D(y', y'')| = D_\mu(y')$. QED.

Note that the Euclidean distance function violates both (h1) and (h2) in all dimensions. The Kullback-Leibler divergence fulfills them if $n = 2$ but violates them if $n > 2$. For $n = 2$, this is easily checked by considering the derivative of the Kullback-Leibler divergence for two dimensions (use D_ε defined in (4.45) and (4.46) for the two-dimensional case instead of D_μ for the arbitrary-dimensional case). A counterexample to fulfillment of (h1) and (h2) for $n = 3$ is given in (4.53).

Now that we have shown that (h1) and (h2) are equivalent, it is easy to show that J_P, L_P and Z_P fulfill the horizon requirement while M_P, R_P and G_P violate it. The following table of derivatives will do (note that $\varepsilon = y - x$ is fixed while x varies and that these derivatives have to be considered with the absolute value for the various degree of confirmation functions in mind). Column 1 is the name of the candidate confirmation function; column 2 is the function with x and $\varepsilon = y - x$ as arguments; column 3 is the derivative $\partial D(x, \varepsilon)/\partial x$ for $|D(x, \varepsilon)| = D(x, \varepsilon)$.

$M_P(x,y)$	ε	0
$R_P(x,y)$	$\log\dfrac{x+\varepsilon}{x}$	$-\dfrac{\varepsilon}{(x+\varepsilon)x}$
$J_P(x,y)$	$\dfrac{x+\varepsilon}{1-x-\varepsilon}-\dfrac{x}{1-x}$	$\dfrac{1}{(1-x-\varepsilon)^2}-\dfrac{1}{(1-x)^2}$
$L_P(x,y)$	$\log\dfrac{(x+\varepsilon)(1-x)}{x(1-x-\varepsilon)}$	(C.3)
$G_P(x,y)$	$\log\dfrac{1-x}{1-x-\varepsilon}$	$\dfrac{h}{(1-x)(1-x-\varepsilon}$
$Z_P(x,y)$	$\dfrac{\varepsilon}{1-x}$ or $\dfrac{\varepsilon}{x}$	$\dfrac{\varepsilon}{(1-x)^2}$ or $-\dfrac{\varepsilon}{x^2}$

with

$$-\frac{((x+\varepsilon)(1-x)-(1-x-\varepsilon)x)(1-2x-\varepsilon)}{(x+\varepsilon)(1-x-\varepsilon)(1-x)x}. \tag{C.3}$$

Appendix D

The Hermitian Form Model

Asymmetric MDS is a promising approach to classify asymmetries in terms of their behaviour. This subsection demonstrates that Chino's asymmetric MDS, both spatial and non-spatial, fails to give us explanations for information theory's violation of TRANSITIVITY OF ASYMMETRY. I am choosing Chino's approach because it is the most general and most promising of all the different asymmetric MDS models (see, for example, Chino and Shiraiwa, 1993, where Chino manages to subsume many of the other approaches into his own).

Multi-dimensional scaling (MDS) visualizes similarity of individuals in datasets. Various techniques are used in information visualization, in particular to display the information contained in a proximity matrix. When the proximity matrix is asymmetrical, we speak of asymmetric MDS. These techniques can be spatial (see for example Chino, 1978), where the proximity relationships are visualized in two-dimensional or higher-dimensional space; or non-spatial (see for example Chino and Shiraiwa, 1993), where the proximity relationships are used to identify data sets with abstract spaces (in Chino's case, finite-dimensional complex Hilbert spaces) and metrics defined on them.

The spatial approach in two dimensions fails right away for information theory because it cannot visualize transitivity violations. The hope for other types of asymmetric MDS is that it would be able to distinguish between well-behaved and ill-behaved asymmetries and either exclude or identify better-behaved candidates than the Kullback-Leibler divergence for measuring the distance between probability distributions. I will use Chino's most sophisticated non-spatial account to show that asymmetric MDS cannot solve this problem. For other asymmetric MDS note that with the Hermitian Form Model Chino seeks to integrate and generalize over all the

other accounts.

Assume a finity proximity matrix. I will work with two examples here to avoid the detailed and abstract account provided by Chino. The first example is

$$D = \begin{bmatrix} 0 & 2 & 3 \\ 3 & 0 & 1 \\ -1 & 2 & 0 \end{bmatrix} \tag{D.1}$$

and allows for easy calculations. The second example corresponds to (4.60), the example for transitivity violation where

$$\hat{D} = \begin{bmatrix} 0.0000 & 0.0566 & 0.0487 \\ 0.0589 & 0.0000 & 0.0499 \\ 0.0437 & 0.0541 & 0.0000 \end{bmatrix}, \tag{D.2}$$

and the elements of the matrix $\hat{d}_{jk} = D_{\text{KL}}(P_j, P_k)$. Note that the diagonal elements are all zero, as no updating is necessary to keep the probability distribution constant.

Chino first defines a symmetric matrix S and a skew-symmetric matrix T corresponding to the proximity matrix such that $D = S + T$.

$$S = \frac{1}{2}(D + D') \text{ and } T = \frac{1}{2}(D - D'). \tag{D.3}$$

Note that D' is the transpose of D, S is a symmetric matrix, and T is a skew-symmetric matrix with $t_{jk} = -t_{kj}$. Next we define the Hermitian matrix

$$H = S + iT, \tag{D.4}$$

where i is the imaginary unit. H is a Hermitian matrix with $h_{jk} = \overline{h_{kj}}$. Hermitian matrices are the complex generalization of real symmetric matrices. They have

special properties (see section 8.9 in Anton and Busby, 2003) which guarantee the existence of a unitary matrix U such that

$$H = U\Lambda U^*, \tag{D.5}$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ with n the dimension of D and λ_k the k-th eigenvalue of H (theorem 8.9.8 in Anton and Busby, 2003). U is the matrix of eigenvectors with the k-th column being the k-th eigenvector. U^* is the conjugate transpose of U. Given example (D.1), the numbers look as follows:

$$H = \frac{1}{2}\begin{bmatrix} 0 & 5-i & 2+4i \\ 5+i & 0 & 3-i \\ 2-4i & 3+i & 0 \end{bmatrix} \tag{D.6}$$

and

$$U = \begin{bmatrix} 0.019 + 0.639i & -0.375 + 0.195i & 0.514 + 0.386i \\ 0.279 - 0.494i & -0.169 - 0.573i & 0.503 + 0.260i \\ -0.519 + 0.000i & 0.681 - 0.000i & 0.516 + 0.000i \end{bmatrix} \tag{D.7}$$

with $\Lambda = \text{diag}(-3.78, 0.0715, 3.71)$. Λ is calculated using the characteristic polynomial $\lambda^3 - 14\lambda + 1$ of H. Notice that the characteristic polynomial is a depressed cubic (the second coefficient is zero), which facilitates computation and will in the end spell the failure of Chino's program for our purposes.

Given example (D.2), the numbers are

$$\hat{H} = \frac{1}{2}\begin{bmatrix} 0.0000 + 0.0000i & 0.0578 - 0.0011i & 0.046 + 0.003i \\ 0.0578 + 0.0011i & 0.0000 + 0.0000i & 0.052 - 0.002i \\ 0.0462 - 0.0025i & 0.0520 + 0.0021i & 0.000 + 0.000i \end{bmatrix} \tag{D.8}$$

and

$$\hat{U} = \begin{bmatrix} 0.351 - 0.467i & -0.543 + 0.170i & -0.578 - 0.006i \\ -0.604 + 0.457i & -0.201 - 0.169i & -0.598 + 0.002i \\ 0.290 - 0.000i & 0.779 + 0.000i & -0.555 + 0.000i \end{bmatrix} \tag{D.9}$$

with $\Lambda = \text{diag}(-0.060, -0.045, 0.104)$.

Chino now elegantly shows how the decomposition of $H = U\Lambda U^*$ defines a seminorm on a vector space. Let $\phi(\zeta, \tau) = \zeta\Lambda\tau^*$. Then (i) $\phi(\zeta_1+\zeta_2, \tau) = \phi(\zeta_1, \tau)+\phi(\zeta_2, \tau)$, (ii) $\phi(a\zeta, \tau) = a\phi(\zeta, \tau)$, and (iii) $\phi(\zeta, \tau) = \overline{\phi(\tau, \zeta)}$. These three conditions characterize an inner product on a finite-dimensional complex Hilbert space, but only if a fourth condition is met: positive (or negative) definiteness ($\phi(\zeta, \zeta) \geq 0$) for all ζ). One might hope that positive definiteness identifies the more well-behaved asymmetries by associating with them a finite-dimensional complex Hilbert space with the norm $\|\zeta\| = \sqrt{\phi(\zeta, \zeta)}$ defined on it (Chino himself speculatively mentioned this hope to me in personal communication).

The hope does not come to fruition. Without a non-trivial self-similarity relation, all seminorms defined as above are indefinite, and thus all cats grey in the night. Not only are well-behaved and ill-behaved asymmetries indistinguishable by the light of this seminorm, even the seminorms for symmetry are indefinite. Not only does this not help our programme, it also puts a serious damper on Chino's, who never mentions the self-similarity requirement (which, given that we are dealing with a proximity matrix, is substantial).

Based on a theorem in linear algebra (see theorem 4.4.12 in Anton and Busby, 2003),

$$\sum_{j=1}^{n} \lambda_j = \text{tr}(A) \tag{D.10}$$

whenever the λ_j are the eigenvalues of A. The reader can easily verify this theorem

by noticing that the roots of the characteristic polynomial add up to the second coefficient (which is the trace of the original matrix). It is well-known that the eigenvalues of a Hermitian matrix are real-valued (theorem 8.9.4 in Anton and Busby, 2003), which is an important component for Chino to define the seminorm $\|\zeta\|$ with the help of ϕ. Unfortunately, using (D.10), the eigenvalues are not only real, but also add up to the trace of H, which is zero unless there is a non-trivial self-similarity relation.

Tversky entertains such self-similarity relations in psychology (see Tversky, 1977, 328), and Chino is primarily interested in applications in psychology. When the eigenvalues add up to zero, however, there will be positive and negative eigenvalues (unless the whole proximity matrix is the null-matrix), which renders the seminorm as defined by Chino indefinite. The Kullback-Leibler divergence is trivial with respect to self-similarity: $D_{\mathrm{KL}}(P, P) = 0$ for all P.

Appendix E

The Powerset Approach Formalized

Let us assume a partition $\{B_i\}_{i=1,\ldots,4n}$ of $A_1 \cup A_2 \cup A_3$ into sets that are of equal measure μ and whose intersection with A_i is either the empty set or the whole set itself (this is the division into rectangles of scenario III). (MAP) dictates that the number of sets covering A_3 equals the number of sets covering $A_1 \cup A_2$. For convenience, we assume the number of sets covering A_1 to be n. Let $\mathcal{C}$, a subset of the powerset of $\{B_i\}_{i=1,\ldots,4n}$, be the collection of sets which agree with the constraint imposed by (HDQ), i.e.

$$C \in \mathcal{C} \text{ iff } C = \{C_j\} \text{ and } t\mu\left(\bigcup C_j \cap A_1\right) = \mu\left(\bigcup C_j \cap A_2\right) \tag{E.1}$$

In figures 7.5 and 7.6 there are diagrams of two elements of the powerset of $\{B_i\}_{i=1,\ldots,4n}$. One of them (figure 7.5) is not a member of $\mathcal{C}$, the other one (figure 7.6) is.

The binomial distribution dictates the expectation EX of X, using simple combinatorics. In this case we require, again for convenience, that n be divisible by t and the 'grain' of the partition A be $s = n/t$. Remember that t is the factor by which (HDQ) indicates that Judy's chance of being in A_2 is greater than being in A_1. In Judy's particular case, $t = 3$ and $\vartheta = 0.75$. We introduce a few variables which later on will help for abbreviation:

$$n = ts \qquad 2m = n \qquad 2j = n - 1 \qquad T = t^2 + 1 \tag{E.2}$$

EX, of course, depends both on the grain of A and the value of t. It makes sense

to make it independent of the grain by letting the grain become increasingly finer and by determining EX as $s \to \infty$. This cannot be done for the binomial distribution, as it is notoriously uncomputable for large numbers (even with a powerful computer things get dicey around $s = 10$). But, equally notorious, the normal distribution provides a good approximation of the binomial distribution and will help us arrive at a formula for G_{pws} (corresponding to G_{ind} and G_{max}), determining the value q_3 dependent on ϑ as suggested by the powerset approach.

First, we express the random variable X by the two independent random variables X_{12} and X_3. X_{12} is the number of partition elements in the randomly chosen C which are either in A_1 or in A_2 (the random variable of the number of partition elements in A_1 and the random variable of the number of partition elements in A_2 are decisively not independent, because they need to obey (HDQ)); X_3 is the number of partition elements in the randomly chosen C which are in A_3. A relatively simple calculation shows that $EX_3 = n$, which is just what we would expect (either the powerset approach or the uniformity approach would give us this result):

$$EX_3 = 2^{-2n} \sum_{i=0}^{2n} i \binom{2n}{i} = n \text{ (use } \binom{n}{k} = \frac{n}{k}\binom{n-1}{k-1}\text{)} \tag{E.3}$$

The expectation of X, X being the random variable expressing the ratio of the number of sets covering A_3 and the number of sets covering $A_1 \cup A_2 \cup A_3$, is

$$EX = \frac{EX_3}{EX_{12} + EX_3} = \frac{n}{EX_{12} + n} \tag{E.4}$$

If we were able to use uniformity and independence, $EX_{12} = n$ and $EX = 1/2$, just as Grove and Halpern suggest (although their uniformity approach is admittedly less crude than the one used here). Will the powerset approach concur with the uniformity approach, will it support the principle of maximum entropy, or will it make another suggestion on how to update the prior probabilities? To answer this question, we must find out what EX_{12} is, for a given value t and $s \to \infty$, using the binomial distribution and its approximation by the normal distribution.

Using combinatorics,

$$EX_{12} = (t+1)\frac{\sum_{i=1}^{s} i\binom{ts}{i}\binom{ts}{ti}}{\sum_{i=0}^{s}\binom{ts}{i}\binom{ts}{ti}} \tag{E.5}$$

Let us call the numerator of this fraction NUM and the denominator DEN. According to the de Moivre-Laplace Theorem,

$$\text{DEN} = \sum_{i=0}^{s}\binom{ts}{i}\binom{ts}{ti} \approx 2^{2n}\sum_{i=0}^{s}\int_{i-\frac{1}{2}}^{i+\frac{1}{2}} \mathcal{N}(\frac{n}{2},\frac{n}{4})(i)\mathcal{N}(\frac{n}{2},\frac{n}{4})(ti)\ di \tag{E.6}$$

where

$$\mathcal{N}(\mu,\sigma^2)(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{E.7}$$

Substitution yields

$$\text{DEN} \approx 2^{2n}\frac{1}{\pi m}\sum_{i=0}^{s}\int_{i-\frac{1}{2}}^{i+\frac{1}{2}} \exp\left(-\frac{(x-m)^2}{m} - \frac{t^2\left(x-\frac{m}{t}\right)^2}{m}\right)\ dx \tag{E.8}$$

Consider briefly the argument of the exponential function:

$$-\frac{(x-m)^2}{m} - \frac{t^2\left(x-\frac{m}{t}\right)^2}{m} = -\frac{t^2}{m}(a''x^2+b''x+c'') = -\frac{t^2}{m}\left(a''(x-h'')^2+k''\right) \tag{E.9}$$

with (the double prime sign corresponds to the simple prime sign for the numerator later on)

$$a'' = \frac{1}{t^2}T \qquad b'' = (-2m)\frac{1}{t^2}(t+1) \qquad c'' = 2m^2\frac{1}{t^2}$$

$$h'' = -b''/2a'' \qquad k'' = a''h''^2 + b''h'' + c'' \tag{E.10}$$

Consequently,

$$\text{DEN} \approx 2^{2n} \exp\left(-\frac{t^2}{m}k''\right) \sqrt{\frac{1}{\pi a'' m t^2}} \int_{-\infty}^{s+\frac{1}{2}} \mathcal{N}\left(h'', \frac{m}{2a''t^2}\right)\, dx \tag{E.11}$$

And, using the error function for the cumulative density function of the normal distribution,

$$\text{DEN} \approx 2^{2n-1} \sqrt{\frac{1}{\pi a'' m t^2}} \exp\left(-\frac{k''t^2}{m}\right) (1 - \text{erf}(w'')) \tag{E.12}$$

with

$$w'' = \frac{t\sqrt{a''}\left(s + \frac{1}{2} - h''\right)}{\sqrt{m}} \tag{E.13}$$

We proceed likewise with the numerator, although the additional factor introduces a small complication:

$$\begin{aligned} \text{NUM} &= \sum_{i=1}^{s} i \binom{ts}{i} \binom{ts}{ti} = \sum_{i=1}^{s} s \binom{ts}{i} \binom{ts-1}{ti-1} \\ &\approx s2^{2n-1} \sum_{i=1}^{s} \mathcal{N}\left(m, \frac{m}{2}\right)(i) \mathcal{N}\left(j, \frac{j}{2}\right)(ti-1) \end{aligned}$$

Again, we substitute and get

$$\text{NUM} \approx s2^{2n-1} \left(\pi\sqrt{mj}\right)^{-1} \sum_{0}^{s-1} \int_{i-\frac{1}{2}}^{i+\frac{1}{2}} \exp\left(a'(x-h')^2 + k'\right)\, dx \tag{E.14}$$

where the argument for the exponential function is

$$-\frac{1}{mj}\left(j(x-m)^2 + mt^2\left(x - \frac{j+1}{t}\right)^2\right) \tag{E.15}$$

and therefore

$$a' = j+mt^2 \qquad b' = 2j(1-m)+2mt\,(t-j) \qquad c' = j(1-m)^2+m\,(t-j-1)^2 \tag{E.16}$$

$$h' = -b'/2a' \qquad k' = a'h'^2 + b'h' + c' \tag{E.17}$$

Using the error function,

$$\text{NUM} \approx 2^{2n-2}\frac{s}{\sqrt{\pi a'}}\exp\left(-\frac{k'}{mj}\right)(1+\text{erf}(w')) \tag{E.18}$$

with

$$w' = \frac{\sqrt{a'}\left(s-\frac{1}{2}-h'\right)}{\sqrt{mj}} \tag{E.19}$$

Combining (E.12) and (E.18),

$$\begin{aligned} EX_{12} &= (t+1)\frac{\text{NUM}}{\text{DEN}} \\ &\approx \frac{1}{2}(t+1)\sqrt{\frac{Tts}{Tts-1}}se^{\alpha_{t,s}} \end{aligned}$$

for large s, because the arguments for the error function w' and w'' escape to positive infinity in both cases (NUM and DEN) so that their ratio goes to 1. The argument for the exponential function is

$$\alpha_{t,s} = -\frac{k'}{mj} + \frac{k''t^2}{m} \tag{E.20}$$

and, for $s \to \infty$, goes to

$$\alpha_t = \frac{1}{2}T^{-2}(2t^3 - 3t^2 + 4t - 5) \tag{E.21}$$

Notice that, for $t \to \infty$, α_t goes to 0 and

$$EX = \frac{n}{EX_{12} + n} \to \frac{2}{3} \tag{E.22}$$

in accordance with intuition T2.