Essays on Learning and Induction

Michael Nielsen

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY
2019

ABSTRACT

Essays on Learning and Induction

Michael Nielsen

What is the correct way to respond to newly acquired information? What methods for updating beliefs and other attitudes are rational? And what makes them rational? This dissertation is a collection of independent essays, each of which addresses these questions. Among other things, I investigate the extent to which Bayesian learning can be considered objective, the circumstances in which rational learning reduces uncertainty and produces consensus, whether rational learning is compatible with disagreement and polarization, and the relationship between long-run and short-run norms for learning.

Contents

Acknowledgments

I would like to thank my advisors, Haim Gaifman and Jessica Collins, for their support and encouragement during the time that I worked on this dissertation. I am also grateful to the other members of my committee—Melissa Fusco, Simon Huttegger, and Achille Varzi—for their helpful feedback during my defense.

My work has benefitted greatly from countless discussions with my fellow graduate students at Columbia. I am especially grateful to Robby Finley, Arthur Heller Britto, Yang Liu, Igancio Ojea Quintana, and Rush Stewart. Thanks also to my coauthors—Rush Stewart and Eric Wofsey—who contributed directly to two of this dissertation's chapters.

I am very grateful to Judith and Isaac Levi for welcoming me into their home early on in my graduate studies and for the many wonderful discussions that we had there.

My biggest thanks goes to my family—my parents, Steve and Cheryl, and my sister, Kayla. I could not have completed this dissertation without their love and support.

Introduction

Here is the central problem in the epistemology of learning as I like to think of it. An agent is about to inquire into some subject of interest to her. Prior to beginning her inquiry, she has various attitudes about the subject at hand—beliefs, preferences, and so on. These attitudes might be based on information gathered in previous inquiries, but they need not be. There are many factors that can influence the attitudes with which an agent begins a new inquiry. The agent may be prepared to make various simplifying assumptions about the problem she is working on in order to make inquiry more tractable. She may be influenced by other agents. For instance, her attitudes may be the outcome of a consensus-seeking procedure designed to facilitate a joint inquiry between her and other members of her community.

And now inquiry begins. The agent gathers new information and the question is, How, if at all, should the agent's initial attitudes change in the light of this new information? Or, put somewhat differently, what methods for converting old attitudes into new ones are good? And what makes them good?

Answers to these question will depend on many things, including the goals of inquiry, the attitudes being updated, the kind of information that inquiry makes available, and so on. To paint a better picture of how an answer might go, let me give an example. The example I'll give is perhaps the most prominent and well-explored account of learning, not only in philosophy, but in statistics and economics as well. This is the *Bayesian* theory of learning.

Bayesian agents begin inquiry with subjective probabilities that represent their levels of confidence in the events of interest to them. Why probabilities? There are several arguments for taking levels of confidence to be probabilities, but the one most relevant to the work below is due to Ramsey (1931) and de Finetti (1974). According to Ramsey's argument, agents reveal their levels of confidence by announcing fair prices for gambles on events of interest. Fair prices are *coherent*

if there's no collection of gambles (a so-called Dutch book) that is certain to net the agent a loss. And, it turns out, fair prices are coherent if and only if they are probabilities. So, in order to be coherent, Bayesian agents begin inquiry with subjective probabilities. On de Finetti's psychologistic approach, these subjective probabilities are hard-wired aspects of the human mind. The culmination of the behavioral approach to understanding subjective probability is due to Savage (1954), for whom probabilities represent uncertainty about a space of acts and outcomes on which an agent has qualitative preferences. It should be mentioned that there are various kinds of Bayesianism that this work will not discuss. Also, in so-called "Bayesian statistics," prior probabilities are often subject to empirical criteria of success, which I will not focus on in what follows.

With a coherent prior in hand, the agent starts to inquire. At each stage of inquiry, it is assumed that agents observe whether or not an event of interest occurs. Agents respond to this information by conditioning their priors; they use Bayes' rule to derive a *posterior* probability from their prior probability and the observed event. Why Bayes' rule? Again, there are several arguments. Most relevant to us is what Bayes' rule delivers. First, it can be shown, under relatively few assumptions, that Bayesian agents are certain that their posteriors will converge to the truth as they observe more and more events. Second, it can be shown that Bayesian agents are certain that joint inquiry will resolve disagreements: provided the priors of two Bayesian agents do not assign probability zero to different events, both agents are certain that their posteriors will merge as shared information accumulates; differences in their priors will "wash out."

In sum, the core of the Bayesian account of learning amounts to the following. Agents update prior subjective probabilities by conditioning on observed events. This method of learning is good because (among, perhaps, other reasons) it guarantees convergence to the truth and washing out of priors as more and more observations are made.[1]

The Bayesian theory will be studied in much more detail in the chapters that follow. I've sketched it here, not only as an example of how one might go about addressing the central questions in the epistemology of learning, but also to note some of the questions that it leaves unanswered. What about attitudes other than subjective probabilities? How should they be updated in the

[1] Although the mathematical results underlying the Bayesian account go back to Lévy (1937), Gaifman and Snir (1982) were the first to offer a detailed interpretation of these results by presenting them as the conjunction of two claims: (I) convergence to the truth and (II) merging of opinions. For further discussion of this point, see the beginning of chapter three.

light of new information? What if inquiry involves more than simply observing events? What if inquiry provides more complicated kinds of information? What if, say, information is conveyed across a "noisy" channel, so that the agent does not observe any particular event with certainty? What mathematical assumptions are required for the convergence results mentioned above? Are they normatively compelling or mere mathematical conveniences? If the latter, what happens when they are relaxed? Why should we care about convergence results in the first place? Why is convergence a laudable feature of learning? These are some of the main questions that will be addressed in what follows.

The dissertation comprises six chapters. They are thematically related—as I hope the remarks above have indicated—but not interdependent in any way. Each chapter is a self-contained essay.

The starting point of Chapter 1 (*Deterministic Convergence and Strong Regularity*) is the idea that Bayesians can counter charges of excessive subjectivity by appealing to convergence results. The idea, which goes back to Savage (1972), was already advertised above: objectionable differences in prior probability judgments will vanish as agents learn, and individual agents will converge to the truth. Glymour (1980), Earman (1992) and others have voiced the complaint that the theorems used to support these claims tell us, not how probabilities updated on evidence will *actually* behave in the limit, but merely how Bayesian agents *believe* they will behave, suggesting that the theorems are too weak to underwrite notions of scientific objectivity and intersubjective agreement. I investigate, in a very general framework, the conditions under which updated probabilities actually converge to a settled opinion and the conditions under which the updated probabilities of two agents actually converge to the same settled opinion. I call this mode of convergence *deterministic*, and derive results that extend those found in Huttegger (2015b). The results here lead to a simple characterization of deterministic convergence for Bayesian learners and give rise to an interesting argument for what I call *strong regularity*, the view that probabilities of non-empty events should be bounded away from zero.[2]

In Chapter 2 (*Persistent Disagreement and Polarization in a Bayesian Setting*), which is joint work with Rush Stewart, we argue against the thesis that rational agents who learn the same evidence are able to resolve their disagreements. In order to do this, we reflect on the significance of

[2] A version of this chapter is forthcoming in *The British Journal for the Philosophy of Science*.

the merging of opinions theorems. A crucial assumption for merging of opinions in the Bayesian setting is that agents' priors assign probability zero to the same events. We argue that this assumption admits neither normative nor descriptive justification. Furthermore, if the assumption is relaxed, even in a very mild way, not only can merging of opinions fail, but opinions can *polarize* despite being updated on an *infinite stream of shared evidence*.[3]

Chapter 3 (*Convergence to the Truth Without Countable Additivity*) connects with a central question in the philosophy of probability: Must (subjective) probabilities be countably additive, or are merely finitely additive probabilities permissible? Countable additivity is a crucial mathematical assumption needed to derive Bayesian convergence theorems. This chapter investigates the conditions under which merely finitely additive Bayesian probabilities converge to the truth. Most of the work in this chapter is of a technical nature, though I argue that the results are relevant to topics currently being discussed by philosophers of science and formal epistemologists.

Chapter 4 (*Speed-Optimal Induction and Dynamic Coherence*) is the most removed from Bayesian theory. In this chapter, which is joint work with Eric Wofsey, we develop new results for frequency prediction that answer problems going back to Reichenbach (1938). A standard way to challenge convergence-based accounts of inductive success, like the one advocated by Reichenbach, is to claim that they are too weak to constrain inductive inferences in the short-term. We respond to such a challenge by building on ideas in Juhl (1994). When it comes to predicting frequencies, we show that speed-optimal convergence—a long-run success condition—induces dynamic coherence in the short-term.

Chapter 5 (*A Short Proof of the Blackwell-Dubins Theorem*) is a short note that provides a simple proof of the Blackwell-Dubins merging of opinions theorem, which plays a central role in earlier chapters of the dissertation.

[3] A version of this chapter is forthcoming in *The British Journal for the Philosophy of Science*.

Chapter 1

Deterministic Convergence and Strong Regularity

1.1 Introduction

Classical Bayesian epistemology is often criticized for being too subjective. Bayesian theory requires only that inquirers have coherent prior probabilities (de Finetti, 1974) and that these probabilities be updated by conditionalizing on evidence. Because it is rationally permissible, on this view, to adopt any probability measure whatsoever as one's prior, the theory allows widespread disagreement between different rational agents' probability judgments.[1] In response to this, the criticism goes, Bayesian theory cannot explain why some probability judgments are rationally superior to others. For example, it cannot explain science's claims to rationality and objectivity.[2]

In response to charges of excessive subjectivity, Leonard Savage and subsequent theorists have shown that, under mild assumptions, the probabilities of two Bayesian agents who learn the same sequence of propositions are *almost sure* to eventually reach a consensus (Savage, 1972, Sections 3.6, 3.7), and they are *almost sure* to converge to the truth. These are the *merging of opinions* and *convergence to the truth* theorems (Section 2), and they have given rise to a great deal of philosophical discussion (Schervish and Seidenfeld, 1990; Earman, 1992; Huttegger, 2015a,b; Weatherson, 2015; Elga, 2016). What they show, essentially, is that as evidence is gathered differences between prior

[1]This is true of classical, subjective Bayesianism in the tradition of de Finetti and Savage. I do not address "objective" varieties of Bayesianism in this essay.

[2]In *The Foundations of Statistics*, Savage describes the objection as follows.

> It is often argued by holders of necessary and objective views alike that...scientific method consists largely, if not exclusively, in finding out what is probably true, by criteria on which all reasonable men agree...Holders of necessary views say that, just as there is no room for dispute as to whether one proposition is logically implied by others, there can be no dispute as to the extent to which one proposition is partially implied by others that are thought of as evidence bearing on it (Savage, 1972, p. 67).

probability assignments are "washed out" and intersubjective agreement is achieved, with probability 1.[3] The theorems purport to show that the widespread disagreement that Bayesianism permits, and which some find objectionable, is a transient phenomenon.[4]

But some dissenters have found the "almost sure" and "with probability 1" qualifications that appear in the merging and convergence results troublesome. They point out, correctly, that these theorems show that Bayesian agents *believe* (with probability 1) that their probabilistic judgments will converge to the truth and merge with those of other Bayesian agents. The results do *not* show that convergence to the truth and merging of opinions *always* occur, only that Bayesian agents *think* they do. This objection has been voiced by Clark Glymour (1980) and John Earman (1992, Chapter 6) and has been developed more recently by Gordon Belot (2013; 2017).[5]

At this juncture, it is natural to ask under what conditions Bayesian learning brings about convergence, not *almost* surely, but *surely*. This paper provides an answer to that question, and it argues that the answer has interesting connections to other philosophical problems. In the remainder of this section I will give an informal summary of the arguments and results that follow.

Our primary object of study is a mode of convergence that I call *deterministic*. When a sequence of probabilities converges deterministically, a limiting probability distribution exists *without qualification*. Deterministic convergence is not accompanied by an "almost sure" hedge. Deterministic convergence has been studied for updates that go by probability kinematics, or Jeffrey conditioning (Skyrms, 1996; Huttegger, 2015b), but we work in a much more general framework and derive results for Bayesian learning as corollaries to the main results. We will show that, under the assumption that learning does not contradict or destroy previously learned information, deterministic convergence is equivalent to a relation called *uniform absolute continuity*. Very roughly, a sequence of updates converges deterministically if and only if there exists a uniform bound on the "amount

[3]Savage's original result is shown under the assumption that agents observe the outcomes of an independent and identically distributed process, such as repeated tosses of a coin with unknown bias. This would make it appear that the unknown bias of the coin gives rise to an objective probability, but Savage remarks (p. 51) that de Finetti's representation theorem for exchangeable processes allows him to avoid appealing to objective probabilities.

[4]Summarizing one such result, Savage and his coauthors write, "This approximate merging of initially divergent opinions is, we think, one reason why empirical research is called 'objective'"(Edwards et al., 1963, p. 197). Even prior to Savage, convergence was taken as a standard for judging probabilistic inference methods (e.g., Reichenbach, 1938, §43).

[5]Haim Gaifman has voiced similar concerns. About the theorems discussed in Section 2, he writes, "'Probability 1' refers of course to the value of our given prior; hence these are coherence results that constitute an inner justification. It is when we come to justify the prior itself that the negative aspects emerge" (Gaifman, 2009, p. 45).

of change" that the updates undergo. This leads to an interesting result about Bayesian learning. Bayesian updates converge deterministically if and only if the prior probabilities of learned propositions are not arbitrarily small. If a Bayesian agent learns a sequence of increasingly "surprising" propositions, then her probabilities will not converge to a stable distribution.

Insofar as deterministic convergence is a desirable outcome of Bayesian learning, the results mentioned above give rise to an argument for a view that I call *strong regularity*. Strong regularity is a strengthening of the view, endorsed by David Lewis (1980), that probabilities of non-empty propositions should be positive (in the literature, Lewis's view is called *regularity*).[6] Strong regularity demands, not only positivity, but also that there exists some positive real number that is strictly less than every probability of a non-empty proposition. As we will see, this is an extremely strong requirement, and one that will probably strike many as unacceptable. This places Bayesians in a difficult position. As there are serious objections to all of the available asymptotic-based responses to complaints about excessive subjectivity, Bayesians may wish to relinquish notions of objectivity altogether and adopt a thoroughgoing subjectivism.

The rest of the paper proceeds as follows. Section 1.2 introduces the mathematical tools that we will need to study deterministic convergence and states the classical convergence to the truth and merging of opinions theorems. In Section 1.3, we present and discuss the main results of the paper. Section 1.4 addresses an objection to the approach taken in Section 1.3. In Section 1.5, we provide an additional characterization result for Bayesian updates and present the argument for strong regularity. Section 1.6 contains concluding remarks.

1.2 Preliminaries

Let Ω be a sample space of elementary events or possible worlds, and let $\mathcal{F}$ denote a sigma-algebra of subsets of Ω.[7] We call members of $\mathcal{F}$ *events* or *propositions*. A standard example used to illustrate these concepts, which we appeal to several times below, is infinite coin tossing. Here, each point in Ω is a countably infinite binary sequence (with 1 representing heads and 0 representing tails, for instance), and $\mathcal{F}$ is the smallest sigma-algebra containing all the events that are determined by

[6] Another name for this view is *Cromwell's rule* (Lindley, 2006).

[7] To say that $\mathcal{F}$ is a sigma-algebra of subsets of Ω means that $\Omega \in \mathcal{F}$ and $\mathcal{F}$ is closed under complementation and countable unions.

finitely many tosses.[8] Included in $\mathcal{F}$ are events like "the first 3 tosses land tails" and also more complicated events like "the limiting relative frequency of heads is 1/2". In general, we call the pair $(\Omega, \mathcal{F})$ a *measurable space*. A *probability measure* on $(\Omega, \mathcal{F})$ is a non-negative, countably additive set function $P : \mathcal{F} \to [0, 1]$ that assigns the sure event Ω probability 1.[9] An event A is said to occur *almost surely* (with respect to P) if $P(A) = 1$. The triple $(\Omega, \mathcal{F}, P)$ is called a *probability space*.

A Bayesian agent is represented by a probability space $(\Omega, \mathcal{F}, P)$, and P is called her *prior*. Bayesian agents update their priors on evidence by a process called *conditionalization*. According to this model of learning, if E is the strongest proposition that an agent learns between two moments in time, then, after learning E, an agent's new probability measure, her *posterior* or her *updated probability*, should be equal to her prior conditional probability $P(\cdot \mid E)$ given E. Mathematically, this can be expressed succinctly by writing

$$P_E(A) = P(A \mid E) := \frac{P(A \cap E)}{P(E)} \quad \text{whenever } P(E) > 0,$$

where A is an arbitrary event, P_E is the agent's posterior after learning E, and the right-most term is simply the definition of conditional probability.

Observe that the conditional probability given E is not defined when $P(E) = 0$. This means that agents who update by conditionalization cannot learn propositions that they previously assigned probability 0. This is an important limitation of Bayesian learning that we will return to below.

In order to explain how the posterior probabilities of Bayesian agents converge to the truth and merge with other Bayesians' posteriors, we must introduce two important assumptions about the evidence that agents learn. These assumptions play an important role not only in this section, but throughout the entire paper. The first assumption is that evidence is *increasing*. The coin tossing example provides a simple instance of increasing evidence. At the beginning of inquiry, the agent is uncertain about which $\omega \in \Omega$ is actual, and this uncertainty is represented by a prior probability. Suppose that the actual sequence of coin tosses is revealed to the agent one toss at a time. That is, at each stage n, the agent learns the first n digits of the actual binary sequence ω. Her evidence

[8] Formally, an event is determined by finitely many tosses if it is of the form $A_1 \times A_2 \times ...$, where each A_i is a subset of $\{0, 1\}$ and $A_i = \{0, 1\}$ for all but finitely many i.

[9] To say that P is countably additive means that $P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$ whenever $\{A_n : n \in \mathbb{N}\}$ is a sequence of pairwise disjoint events in $\mathcal{F}$.

is increasing in the sense that the information learned at stage n is still available at stage $n+1$.

To explain what it means for evidence to be increasing with more precision and generality we need a few more mathematical concepts. The framework that we will adopt to explain these concepts is essentially the same as that of Kalai and Lehrer (1994) and Huttegger (2015b). A *sigma-subalgebra* of $\mathcal{F}$ is a subset of $\mathcal{F}$ that is itself a sigma-algebra of subsets of Ω. Intuitively, sigma-algebras (and sigma-subalgebras) represent bodies of information. Returning to the example of infinite coin tossing, the sigma-subalgebra

$$\mathcal{F}_1 = \big\{\emptyset, \{\omega \in \Omega : \omega \text{ begins with } 1\}, \{\omega \in \Omega : \omega \text{ begins with } 0\}, \Omega\big\}$$

represents the information corresponding to the first coin toss. $\mathcal{F}_1$ is a particularly simple sigma-subalgebra because it is generated by a finite partition of Ω. That is, $\mathcal{F}_1$ is the smallest sigma-algebra containing the partition that separates sequences that begin with 0 from sequences that begin with 1. Consider now a refinement of this partition, namely the one that separates sequences according to the first two coin tosses

$$\begin{aligned}\big\{&\{\omega \in \Omega : \omega \text{ begins with } 11\}, \{\omega \in \Omega : \omega \text{ begins with } 10\},\\ &\{\omega \in \Omega : \omega \text{ begins with } 01\}, \{\omega \in \Omega : \omega \text{ begins with } 00\}\big\}.\end{aligned} \tag{1.1}$$

If we let $\mathcal{F}_2$ be the smallest sigma-algebra containing the partition in (1.1), then $\mathcal{F}_2$ represents the information corresponding to the first two coin tosses.[10] Notice that $\mathcal{F}_1 \subseteq \mathcal{F}_2$, and if we define $\mathcal{F}_n$ to be the sigma-subalgebra that represents the information corresponding to the first n coin tosses, following the pattern above, then we obtain an increasing sequence $\{\mathcal{F}_n : n \in \mathbb{N}\}$ of sigma-subalgebras of $\mathcal{F}$. Such a sequence is called a *filtration*. Filtrations provide a natural means of representing increasing information or evidence.

In general, when we assume that evidence is increasing, we are assuming the existence of a filtration. In our coin tossing example, each $\mathcal{F}_n$ is generated by a finite partition, namely the partition that separates binary sequences according to the first n tosses. But, in general, a sigma-subalgebra need not be generated by a partition, so filtrations can have a much more complicated

[10] In general, if $\mathcal{E}$ is a finite partition of Ω, then the smallest sigma-algebra containing $\mathcal{E}$ is the collection of all unions of elements of $\mathcal{E}$ together with the empty set.

structure than the example that we have been considering. Nonetheless, it is helpful to think of a filtration as corresponding to a sequence of finer and finer partitions of the sample space, as in the coin tossing case. For this reason, in the main text we will assume that for any filtration $\{\mathcal{F}_n : n \in \mathbb{N}\}$, each $\mathcal{F}_n$ is generated by a finite partition of Ω. It bears emphasizing that we make this assumption simply in order to ease the exposition. The results discussed in this section and the main result in Section 1.3 (Theorem 1.2) do not depend on it.[11]

Our second assumption is that increasing evidence is *complete*. Very roughly, this means that the evidence eventually informs the agent about every event of interest to her. Formally, we assume that the filtration $\{\mathcal{F}_n : n \in \mathbb{N}\}$ *generates* $\mathcal{F}$. This means that $\mathcal{F}$ is the smallest sigma-algebra containing $\bigcup_n \mathcal{F}_n$. What this amounts to, intuitively, is that every event in $\mathcal{F}$ can be approximated to an arbitrary degree of precision by events that the agent learns.[12] In the coin tossing example, the assumption of complete evidence means that the agent's prior cannot be defined on events concerning, say, the denomination of the coin being tossed. Since the filtration in this example encodes information only about the outcome of tosses, there is no way to generate events like "the coin is a nickel" from this filtration.

Assuming that evidence is increasing and complete, we can now state the convergence to the truth theorem. First, the truth about a proposition A depends on which world $\omega \in \Omega$ is the actual one. We can represent this using the *indicator function* $\mathbf{1}_A$ *of* A defined by

$$\mathbf{1}_A(\omega) = \begin{cases} 1, \text{ if } \omega \in A \\ 0, \text{ if } \omega \notin A. \end{cases}$$

The indicator function returns 1 if the proposition A is true at the world ω and 0 if A is false at ω. Finally, let $E_n(\omega)$ denote the cell of the partition that generates $\mathcal{F}_n$ which contains ω. Intuitively, $E_n(\omega)$ represents the event that a Bayesian agent conditionalizes on at stage n if ω is the actual world. In the coin tossing example, $E_n(\omega)$ is the set of all binary sequences whose first n digits are the same as ω's. If the first toss lands heads in ω, for instance, then $E_1(\omega) = \{\omega \in \Omega : \omega \text{ begins with } 1\}$. The *convergence to the truth theorem* says that for all $A \in \mathcal{F}$,

[11] See footnote 15 for more on this point.

[12] Once again, the assumption that evidence is complete is not strictly necessary for stating the results below, but it eases the exposition considerably. See footnote 15.

$$P\big(\{\omega \in \Omega : \lim_{n\to\infty} P(A \mid E_n(\omega)) = \mathbf{1}_A(\omega)\}\big) = 1.^{13} \tag{1.2}$$

In other words, for all events A, let

$$C_A = \Big\{\omega \in \Omega : \lim_{n\to\infty} P(A \mid E_n(\omega)) = \mathbf{1}_A(\omega)\Big\}$$

be the event that a Bayesian agent, with prior P, has posterior probabilities for A that converge to the truth about A. Then, the convergence to the truth theorem says that $P(C_A) = 1$ for all events A. Convergence to the truth about A occurs almost surely. It should be emphasized that (1.2) is consistent with the existence of events A and possible worlds ω for which the sequence $\{P(A \mid E_n(\omega)) : n \in \mathbb{N}\}$ of posteriors does *not* converge to the truth about A. Failure to converge to the truth, in this sense, is not ruled out by the convergence to the truth theorem. What the theorem says is that Bayesians are obliged to assign probability 0 to the set of worlds for which convergence to the truth about A fails.

The convergence to the truth theorem tells us about the behavior of a *single* Bayesian agent's posterior probabilities with respect to a *fixed* event A. The merging of opinions theorem, to which we now turn, tells us about the behavior of *two* Bayesian agents' *overall* posterior probability measures. Roughly speaking, the merging of opinions theorem states that if two Bayesian agents' priors assign probability 0 to the same events, and if they conditionalize on the same increasing and complete evidence, then they will almost surely reach a consensus in the limit.[14]

To explain the theorem more precisely, let P and Q be probabilities on the same measurable space $(\Omega, \mathcal{F})$. It is natural to define the distance between P and Q with respect to the event A as $|P(A) - Q(A)|$. To extend this to a notion of *overall* distance—one that is independent of a particular event A—we define the *total variation distance* d between two probabilities P and Q to be the *supremum*, or least upper bound (in finite cases, the maximum), of their distances with

[13]Note that if $P(E_n(\omega)) = 0$ for some n and ω, then the conditional probability $P(A \mid E_n(\omega))$ can be defined arbitrarily because convergence to the truth is allowed to fail on a set of probability 0. (1.2) is sometimes called the Levy 0-1 Law (Durrett, 2010, Section 4.5, Theorem 5.8). See Schervish and Seidenfeld (1990) for further discussion of this result.

[14]The result is originally due to Blackwell and Dubins (1962) and was discovered independently by Gaifman and Snir (1982) (Theorem 2.1, II). Although the general framework of Gaifman and Snir is quite different from the one adopted here, it is straightforward to translate the proof of their Theorem 2.1 into the framework of this paper.

respect to events:

$$d(P, Q) := \sup_{A \in \mathcal{F}} |P(A) - Q(A)|.$$

Note that P and Q are identical probability measures if and only if $d(P, Q) = 0$.

As before, we assume that evidence is increasing and complete. We also need to assume that P and Q do not differ too dramatically in their probability assignments. In particular, we require that P and Q be *mutually absolutely continuous*: they assign probability 0 (and, hence, probability 1) to exactly the same events. This is a natural requirement, especially in the Bayesian context. If $P(A) = 0$ then no amount of conditionalizing can raise the P-posterior probability of A above 0. If the Q-posterior probabilities of A tend to some positive value, then P and Q cannot achieve agreement in the long run. For all $\omega \in \Omega$, let $P_n(\omega)$ be the posterior probability for P at stage n, i.e. $P_n(\omega) = P(\cdot \mid E_n(\omega))$, and similarly for $Q_n(\omega)$. Under the assumption of mutual absolute continuity, the *merging of opinions theorem* states that

$$P\big(\{\omega \in \Omega : \lim_{n\to\infty} d(P_n(\omega), Q_n(\omega)) = 0\}\big) = 1 = Q\big(\{\omega \in \Omega : \lim_{n\to\infty} d(P_n(\omega), Q_n(\omega)) = 0\}\big).^{15} \quad (1.3)$$

The merging of opinions theorem tells us that both P and Q are certain (with probability 1) that their posterior probability assignments will become arbitrarily close to each other as the evidence

[15] As noted in the main text, neither (1.2) nor (1.3) depend on the assumption that each $\mathcal{F}_n$ in the filtration $\{\mathcal{F}_n : n \in \mathbb{N}\}$ is generated by a finite partition. Also, the assumption that $\{\mathcal{F}_n : n \in \mathbb{N}\}$ generates $\mathcal{F}$ can be relaxed. In this more general setting, we must appeal to conditional probabilities $P_n = P(\cdot \mid \mathcal{F}_n)$ given sigma-subalgebras $\mathcal{F}_n$. The existence of these objects is non-trivial but is guaranteed by the Radon-Nikodym theorem. Uniqueness, on the other hand, is guaranteed only with probability 1. Hence, P_n may have various *versions* that differ from each other on a set of P-probability 0. Conditional probabilities given sigma-subalgebras, despite their name, are not guaranteed to obey the probability axioms in all possible worlds. For a summary of some of the issues involved see Seidenfeld (2001) or Huttegger (2015b). Now, with $\mathcal{F}_\infty$ the sigma-algebra generated by $\{\mathcal{F}_n : n \in \mathbb{N}\}$, (1.2) generalizes to

$$P\big(\{\omega \in \Omega : P_n(A)(\omega) \to \mathbf{1}_A(\omega)\}\big) = 1$$

for all $A \in \mathcal{F}_\infty$. Under the additional assumption that P_n and Q_n are *regular conditional distributions*, or, in Blackwell and Dubins's terminology *predictive* conditional probabilities, (1.3) holds as stated with $d(P, Q) := \sup_{A \in \mathcal{F}_\infty} |P(A) - Q(A)|$. We note that because the members of the filtration in the main text are assumed to be generated by finite partitions, it follows that there are versions of the conditional probabilities in (1.3) that are regular conditional distributions. Finally, it should be noted that both the convergence to the truth theorem and merging of opinions theorem depend crucially on the assumption that P and Q are countably additive. On the topic of convergence to the truth for finitely additive probabilities, see Zabell (2002) for discussion and references. More recently, Pomatto et al. (2014) and Pomatto and Sandroni (2018) have investigated merging of opinions in a finitely additive context.

accumulates.[16] In the limit, P and Q will achieve consensus (almost surely). As with convergence to the truth, it is important to emphasize that (1.3) is a statement about the prior probability judgments of P and Q. There may be possible worlds ω for which the sequence $\{d(P_n(\omega), Q_n(\omega)) : n \in \mathbb{N}\}$ does *not* approach 0. What (1.3) says is that mutually absolutely continuous Bayesian agents must assign probability 0 to this set of worlds.

I close this section with some brief remarks about absolute continuity. We say that Q is *absolutely continuous* with respect to P and write $Q \ll P$ if for all events A we have $Q(A) = 0$ whenever $P(A) = 0$. Two probabilities are mutually absolutely continuous when $P \ll Q$ and $Q \ll P$ (this is equivalent to the definition given above). If Q comes from P by conditionalization on the event E, then $Q \ll P$ because $Q(A) = P(A \mid E) = 0$ whenever $P(A) = 0$. Absolute continuity plays an important role in the results that follow.

1.3 Deterministic Convergence

In response to Bayesians' claims that the convergence and merging results just discussed can counter charges of excessive subjectivity, John Earman has voiced the following complaint.

> Some of the prima facie impressiveness of these results disappears in the light of their narcissistic character, i.e. the fact that the notion of 'almost surely' is judged by $[P]$...'almost surely' sometimes serves as a rug under which some unpleasant facts are swept (Earman, 1992, p. 147–148).

The fact is that Bayesian learning does not guarantee convergence to the truth, or convergence to a consensus, in a wide range of learning scenarios. What the convergence theorems demonstrate is that Bayesian agents are compelled to assign probability zero to these scenarios, no matter how great their extent. Glymour writes,

> The [convergence to the truth] theorem does not tell us that in the limit any rational Bayesian will assign probability 1 to the true hypothesis and probability 0 to the rest;

[16] Schervish and Seidenfeld (1990) extend the Blackwell and Dubins theorem, showing results in which sets of conditionalized probabilities merge uniformly. Stewart and Nielsen (2019) generalize this a bit, showing a result in which a set of probabilities updated by probability kinematics merges uniformly.

> it only tells us that rational Bayesians are certain that [they] will (Glymour, 1980, p. 73).[17]

In view of Earman's and Glymour's objections, it is natural to ask under what conditions the troublesome almost surely qualifications can be removed from the Bayesian convergence results. Our aim is to provide an answer to this question.

To that end, we will depart somewhat from the model of Bayesian learning that was explained in the previous section. In particular, we will no longer seek results about an agent's probabilistic *beliefs* about the behavior of her updated probabilities. Rather, we want to know under what conditions *actual* sequences of updates converge to a limit. It is my sense that this notion of convergence, which I shall call *deterministic convergence*, has not received sufficient attention in the philosophical literature. There are some precursors to the present work, however, and before proceeding with further details I would like to contrast my approach with the others'.

One way to study the conditions under which Bayesian learning guarantees convergence is to describe the events in which convergence occurs or fails to occur using non-probabilistic concepts. This approach depends on the underlying measurable space $(\Omega, \mathcal{F})$ having some additional structure. The relevant mathematical structure is a *topology*. Using topological concepts, Gordon Belot, building on work in Kelly (1996), has developed Earman's objection into a powerful critique of Bayesian learning. His argument turns on the observation that, although Bayesian agents assign probability zero to the event that they fail to converge to the truth, this event is, in some cases, "very large" or "typical" in a topological sense. This leads him to conclude that the convergence theorems "constitute a real liability for Bayesianism by forbidding a reasonable epistemological modesty" (Belot, 2013).[18] Belot's critique is that Bayesianism mandates immodest probabilistic certainty in convergence, despite the fact that convergence can fail to occur in typical sets of possible worlds.

This argument, while interesting, depends on some assumptions that I do not wish to make. In particular, it assumes that topological structure is somehow relevant to epistemological questions

[17]Similar points have been raised by Kelly et al. (1997) and Howson (2000). A related point arises in the context of calibration. As Dawid (1982) shows, Bayesians forecasts are well calibrated with probability 1, but "In practice...it is rare for probability forecasts to be well calibrated...and no realistic forecaster would believe too strongly in his own calibration performance."(p. 608).

[18]The argument also appears in Belot (2017).

about modesty and reasonable belief. This strikes me as not obvious and even implausible, but this issue is outside the scope of the present paper. For discussion of this point, I refer the reader to Huttegger (2015a), Elga (2016), and Cisewski et al. (2018). My aim is to study the *probabilistic* conditions under which Bayesian learning guarantees convergence, not the topological ones. There are some results on this topic (Skyrms, 1996; Huttegger, 2015b), which will be discussed below. But now it is time to develop the framework of deterministic convergence.

From now on, we will use a very general notion of probabilistic learning. As above, we let $(\Omega, \mathcal{F}, P)$ be a fixed probability space, and we interpret P as an agent's prior. A probability measure P' on $(\Omega, \mathcal{F})$ is an *update* of P if P' is absolutely continuous with respect to P (recall that this means P' assigns probability zero to an event whenever P does). Similarly, a sequence $\{P_n : n \in \mathbb{N}\}$ of probability measures on $(\Omega, \mathcal{F})$ is a *sequence of updates* of P if, for all $n \in \mathbb{N}$, P_n is absolutely continuous with respect to P.[19] As we remarked in Section 2, updates by conditionalization are absolutely continuous, so Bayesian learning is a special case of updating in the current sense. In this section, when we say that a sequence $\{P_n : n \in \mathbb{N}\}$ of updates of P *goes by conditionalization* or *is a sequence of conditionalizations of* P, we mean that there exists a decreasing sequence $\{E_n : n \in \mathbb{N}\}$ of events with positive P-probability and $P_n = P(\cdot \mid E_n)$ for all n.[20]

Our notion of convergence will be very general as well. We say that a sequence $\{P_n : n \in \mathbb{N}\}$ of updates of P *converges deterministically* if

$$\lim_{n\to\infty} P_n(A) \text{ exists for all } A \in \mathcal{F}.^{21}$$

If the sequence $\{P_n : n \in \mathbb{N}\}$ of updates of P converges deterministically, then the *deterministic limit* P_∞ of this sequence is the set function defined by $P_\infty(A) = \lim_{n\to\infty} P_n(A)$ for all events A. Note that the definition of deterministic convergence does not guarantee *a priori* that P_∞ is a probability measure. To say that a sequence of updates of P converges deterministically means that, for each event A, the values of $P_n(A)$ eventually settle down to a limit as n gets very large.

[19] It may be natural to require that a sequence of updates of P satisfy $P_n \ll P_{n-1}$ for all $n \in \mathbb{N}$. Sequences of conditionalizations satisfy this condition, for example. However, as our results do not require this additional assumption, we do not impose it in what follows.

[20] To say that the sequence $\{E_n : n \in \mathbb{N}\}$ is decreasing means that $E_1 \supseteq E_2 \supseteq \ldots$.

[21] In mathematics, this mode of convergence is called *setwise*. Skyrms (1996) has called this convergence to a "maximally informed opinion." Our terminology is meant to highlight the contrast with almost sure convergence.

Deterministic convergence rules out, for instance, the possibility that $P_n(A)$ oscillates between 0.4 and 0.6 forever. Unlike the classical Bayesian convergence theorems, we do not restrict ourselves to studying the conditions under which convergence is to the *truth*. Rather, we are interested in the more general phenomenon in which updates approach *some* stable limit.

A few remarks are in order about the generality of our framework, which some may find objectionable. First, some philosophers may be interested only in the special case in which convergence is to the truth. But the results to come should still be of interest to those philosophers because convergence to the truth is a special case of deterministic convergence: in order to converge to the truth, one's probabilities cannot oscillate forever. Foreshadowing a bit, the results below have a negative character, showing that deterministic convergence is quite difficult to achieve. As the limitations of deterministic convergence are also limitations of convergence to the truth—the latter being a special case of the former—the arguments below can be applied directly to convergence to the truth. Second, it may be objected that our notion of update is too general to be of interest: one should study particular update rules (like conditionalization) on a case by case basis. In response to this, note that we will derive results for Bayesian learning as corollaries of the main theorems. Also, as there is no principled upper bound on the number of alternatives to Bayesian learning that philosophers are likely to propose (there are many in the literature already[22]), it is useful to study a large family of potential alternatives in a general framework like ours.

We can now begin working towards our main results. Simon Huttegger (2015b) has provided two conditions that are sufficient to guarantee deterministic convergence under a modest generalization of Bayesian learning called probability kinematics, or Jeffrey conditioning (Jeffrey, 1992). After introducing these conditions, we will see that they can be used to completely characterize deterministic convergence in our more general framework.

In order to introduce the first condition, we follow the classical results in assuming that evidence is increasing and complete. As explained in the previous section, this assumption is formalized by equipping $(\Omega, \mathcal{F})$ with a filtration $\{\mathcal{F}_n : n \in \mathbb{N}\}$ that generates $\mathcal{F}$. In the previous section, since we were discussing conditionalization exclusively, we assumed that at each stage n, the agent learns

[22]These alternative update procedures include Jeffrey conditioning (or probability kinematics) (Jeffrey, 1992) and various parameterizations thereof (Field, 1978; Wagner, 2002), imaging (Lewis, 1976; Leitgeb, 2017), and Gallow's rule for learning theory-dependent evidence (Gallow, 2014).

with certainty (with probability 1) which member of $\mathcal{F}_n$ is true. But as proponents of probability kinematics have long pointed out, "learning experiences need not be like that at all" (Huttegger, 2015b, p. 613). In the coin tossing case, for instance, information about the first n tosses could be conveyed across a noisy channel. This sort of "uncertain evidence" may not justify certainty about the outcome of the first n tosses, but it may prompt some sort of probabilistic update. For example, if a noisy but fairly reliable channel conveys the information that the first toss lands hands, it may be reasonable for one to update one's prior so that the posterior probability of $\{\omega \in \Omega : \omega \text{ begins with } 1\}$ is high but not quite 1. In this section, we now allow for these sorts of learning experiences, generalizing the frameworks of Bayesian conditionalizing and probability kinematics.[23]

Our first condition requires sequences of updates based on potentially uncertain evidence to exhibit a certain kind of stability. Basically, the requirement is that later learning experiences do not contradict earlier ones. More precisely, the condition states that the probabilities assigned to a partition at stage n are unchanged at time $n+1$. Mathematically, we have

$$P_{n+1}(F) = P_n(F) \text{ for all } n \in \mathbb{N} \text{ and all } F \in \mathcal{F}_n.^{24} \tag{M}$$

Note that sequences of conditionalizations always satisfy (M). On the Bayesian model of learning, when a proposition is learned at time n, it is assigned probability 1, and its probability cannot decrease from 1 at later times.[25]

(M) requires that learned information is never lost or changed. In learning scenarios in which (M) fails, an agent's probabilities for some event $F \in \mathcal{F}_n$ are free to oscillate indefinitely, and she need not converge. Brian Skyrms (1996) has argued that (M) is a consequence of diachronic coherence. And Simon Huttegger (2014; 2015b) has argued that (M) is a constraint on rational or "genuine" learning. For my part, I am not convinced that (M) has a distinguished normative or metaphysical status (see the remarks at the end of this section) and wish to remain neutral about

[23]Gyenis and Rédei (2017) also discuss a model of probabilistic learning that generalizes conditionalization and probability kinematics. Their model is not as general as ours, however. The reader should keep in mind that the only constraint on updates in our model is absolute continuity with respect to the prior.

[24]Note that (M) implies that $P_m(F) = P_n(F)$ for all $m \geq n$ and $F \in \mathcal{F}_n$.

[25]Some have found this feature of Bayesian learning objectionable (Levi, 1980; Jeffrey, 1992; Williamson, 2002).

this issue in the present paper.

The second condition is somewhat more complicated. Let $\{P_n : n \in \mathbb{N}\}$ be a sequence of updates of $(\Omega, \mathcal{F}, P)$. We say that this sequence is *uniformly absolutely continuous* with respect to P if for all $\epsilon > 0$, there exists a $\delta > 0$ such that for all $n \in \mathbb{N}$ and all $A \in \mathcal{F}$ we have

$$P_n(A) < \epsilon \text{ whenever } P(A) < \delta.$$

Intuitively, we can think of uniform absolute continuity as requiring a uniform bound on the amount of change that P undergoes in updating to P_n. The change cannot be arbitrarily drastic. Put another way, still roughly, uniform absolute continuity *rules out* learning scenarios in which prior probabilities are very small and the corresponding posterior probabilities are very large. In order to explain this new property and make the previous comments more precise, it is useful to consider an example in which uniform absolute continuity fails.

Example 1. Let $(\Omega, \mathcal{F}, P)$ be a countable probability space defined by $P(\{\omega_n\}) = 2^{-n}$ for all $n \in \mathbb{N}$.[26] We let the sequence $\{P_n : n \in \mathbb{N}\}$ of updates be given by conditionalization on the events $E_n = \Omega - \bigcup_{i=1}^{n} \{\omega_i\}$. Since P assigns positive probability to every non-empty event, each P_n is absolutely continuous with respect to P. In order to show that $\{P_n : n \in \mathbb{N}\}$ is not *uniformly* absolutely continuous with respect to P, we must show that there exists an $\epsilon > 0$ such that for all $\delta > 0$ there exists an $n \in \mathbb{N}$ and an $A \in \mathcal{F}$ with $P(A) < \delta$ and $P_n(A) \geq \epsilon$.

To that end, let $\epsilon = 1/2$ and let δ be arbitrary. Notice that $P(E_n) = 2^{-n}$, which is arbitrarily close to 0 for large n. So we can choose n sufficiently large so that $P(E_n) < \delta$. But, as P_n comes from conditionalizing P on E_n, we have $P_n(E_n) = 1 \geq 1/2$. The key feature of this example, which we explore further below, is that the prior probabilities of the learned events E_n are arbitrarily small. △

Although the definition of uniform absolute continuity applies to arbitrary sequences of updates, later on we will see (Section 1.5) that the property admits a simple characterization when the updates are conditionalizations.

Huttegger (2015b, Theorem 7.1) shows that deterministic convergence is achieved under the

[26] Throughout the paper, whenever Ω is countable it is assumed that $\mathcal{F}$ is the powerset of Ω.

assumptions that updating is by probability kinematics (a special case of the present framework), that updates are uniformly absolutely continuous with respect to their prior, and that (M) holds.[27] It is natural to inquire after a converse. In particular, what role does the technical-looking uniform absolute continuity property play in Huttegger's result? Our first theorem shows that, somewhat surprisingly, uniform absolute continuity is a necessary condition for deterministic convergence.

Theorem 1.1. *Suppose that the sequence $\{P_n : n \in \mathbb{N}\}$ of updates of P converges deterministically to the set function P_∞. Then $\{P_n : n \in \mathbb{N}\}$ is uniformly absolutely continuous with respect to P and P_∞ is a probability measure on $(\Omega, \mathcal{F})$ with $P_\infty \ll P$.*

Proof. This is an immediate consequence of the Vitali-Hahn-Saks Theorem, a deep result of measure theory (see the Appendix, Theorem 1.7). To summarize, that result guarantees the desired uniform absolute continuity of $\{P_n : n \in \mathbb{N}\}$ with respect to P. It also guarantees that the deterministic limit P_∞ is a finite measure on $(\Omega, \mathcal{F})$ with $P_\infty \ll P$. To finish, it remains to verify that P_∞ is a probability measure. Simply note that $\{P_n(\Omega) : n \in \mathbb{N}\}$ is a constant sequence, so by deterministic convergence we have

$$1 = P_n(\Omega) \to P_\infty(\Omega) = 1$$

as $n \to \infty$. The rightmost equality is all that needs to be shown, so we can conclude. □

Not only does a converse of Huttegger's theorem hold in our more general setting, it is also possible to prove a generalization of his result, which leads to the characterization of deterministic convergence that was advertised above. The next theorem states that (M) and uniform absolute continuity are sufficient to guarantee deterministic convergence for *any* sequence of updates (we needn't assume with Huttegger that updating is by probability kinematics).

Theorem 1.2. *Let $\{P_n : n \in \mathbb{N}\}$ be a sequence of updates of P that satisfies (M). If $\{P_n : n \in \mathbb{N}\}$ is uniformly absolutely continuous with respect to P, then P_n converges deterministically to a limit P_∞, and P_∞ is a probability measure on $(\Omega, \mathcal{F})$ with $P_\infty \ll P$.*

All of the proofs not stated in the main text can be found in the Appendix.[28]

[27] We remark that Huttegger shows deterministic convergence, but does not argue that the limiting set function P_∞ is a probability measure. In fact, it is, as we point out below. This is not merely a technical point: it would be disappointing to achieve convergence only to discover that one's limiting distribution is incoherent.

[28] The conclusion of Theorem 1.2 is strengthened by Theorem 1.8, which is also in the Appendix.

Combining the last two results we have a characterization of deterministic convergence in terms of uniform absolute continuity and (M).

Theorem 1.3. *Let $\{P_n : n \in \mathbb{N}\}$ be a sequence of updates of P that satisfies (M). Then $\{P_n : n \in \mathbb{N}\}$ converges deterministically to a probability measure P_∞ on $(\Omega, \mathcal{F})$ if and only if the sequence $\{P_n : n \in \mathbb{N}\}$ is uniformly absolutely continuous with respect to P.*

Theorem 1.3 indicates that uniform absolute continuity is more than a mere technical device. Rather, it is somehow essential to deterministic convergence, even in our very general setting. To shed more light on the situation, we note that the last result yields as an immediate corollary a simple characterization of deterministic convergence for Bayesian learning by conditionalization.

Corollary 1.1. *Let $\{P_n : n \in \mathbb{N}\}$ be a sequence of conditionalizations of P. Then $\{P_n : n \in \mathbb{N}\}$ converges deterministically to a probability measure P_∞ on $(\Omega, \mathcal{F})$ if and only if the sequence $\{P_n : n \in \mathbb{N}\}$ is uniformly absolutely continuous with respect to P.*

Proof. As discussed above, if $\{P_n : n \in \mathbb{N}\}$ is a sequence of conditionalizations of P, then it is a sequence of updates of P that satisfies (M), and the result follows from Theorem 1.3. □

Example 2. In Example 1, we exhibited a sequence $\{P_n : n \in \mathbb{N}\}$ of conditionalizations of P that is not uniformly absolutely continuous with respect to P. Corollary 1.1 indicates that this sequence does not converge deterministically. Let us exhibit an event A for which the limit of the sequence $\{P_n(A) : n \in \mathbb{N}\}$ does not exist. Note that the failure of deterministic convergence does not imply that the limit of $\{P_n(A) : n \in \mathbb{N}\}$ does not exist for all events A. For example, if A is a finite set, then for large enough n, $A \cap E_n = \emptyset$. From this it follows that the limit of $\{P_n(A) : n \in \mathbb{N}\}$ exists and is equal to 0.

Consider, now, the event A_0 that consists of all even-indexed ω_n. That is, let $A_0 = \{\omega_n : n \text{ even}\}$. After some straightforward calculations, we find that $P_n(A_0) = 1/3$ if n is even and $P_n(A_0) = 2/3$ if n is odd.[29] In this example, the probabilities $P_n(A_0)$ oscillate forever and never reach a stable limit. △

[29]For n even, we have $P(A_0 \cap E_n) = (1/2^{n+2})/(3/4)$. Hence, $P_n(A_0) = P(A_0 \cap E_n)2^n = 1/3$. For n odd, we have $P(A_0 \cap E_n) = (1/2^{n+1})/(3/4)$ and $P_n(A_0) = 2/3$.

I conclude this section with some remarks about the conditions that appear in our results. I do not claim that these conditions are normative, nor that they are constitutive of "genuine learning." My aim, as stated at the beginning of this section, has been to find relatively simple conditions that are characteristic of deterministic convergence. Insofar as a theory of rational probabilistic learning aims to secure deterministic convergence, it must impose the conditions that appear in the theorems above. Insofar as one's normative standards judge such an imposition too severe, one's theory of learning cannot make use of the notion of deterministic convergence. For Bayesians in this latter group, the problem remains to provide a compelling response to objections about excessive subjectivity. In Section 1.5, we will see that further problems arise for Bayesians in the former group as well.

1.4 Consensus in the Limit

Before discussing those problems, however, I would like to address a potential objection to the approach taken in the previous section. Our study of the asymptotics of probabilistic learning was motivated by the charge that Bayesianism is too subjective. The original, simple idea of Savage and his followers was that this charge can be countered by showing that the disagreements between different agents' probability judgements eventually vanish when the agents update on shared evidence. But, the present objection goes, we have not discussed anything like Savage's idea in the present framework. We have only studied the conditions under which an *individual* agent converges to a stable limiting probability distribution. This leaves open the possibility that different agents converge to *different* limiting distributions. If that were to occur, objectionable disagreements would persist indefinitely, and, some might object, that would indicate that the framework of deterministic convergence is not suitable for studying the critiques of Bayesianism that motivated us at the outset.

Given some mild assumptions about what it means for two agents to learn the same evidence, it can be shown that different agents converge deterministically to the *same* limiting distribution. This result should allay the worry that the framework of deterministic convergence is not well-suited to address the issues with which Savage and his followers are concerned. I do not claim that other, more serious worries about deterministic convergence are answered here. We will be turning

to some of those in the next section.

We will assume that two agents learn the same evidence if for each n, their updated probabilities of events in $\mathcal{F}_n$ are the same. Let P and Q be the probabilities of two distinct agents, defined on the same measurable space, and let $\{P_n : n \in \mathbb{N}\}$ and $\{Q_n : n \in \mathbb{N}\}$ be sequences of updates of P and Q, respectively. Formally, we require that

$$P_n(A) = Q_n(A) \text{ for all } n \in \mathbb{N} \text{ and all } A \in \mathcal{F}_n,$$

and we say that P and Q *learn the same evidence*. This is a natural way to formalize the notion of learning the same evidence in the current framework. At each stage n, both agents receive the same (potentially "noisy" or "uncertain") evidence and they agree about that evidence's impact on their posterior probability assignments. The case in which two agents conditionalize on the same event is a special case of the current definition. There are, to be sure, other ways of understanding learning the same evidence, but we will not enter that discussion here. We refer the reader to Huttegger (2015b) and Wagner (2002, 2003).

The next result states that, under the same assumptions that we made in the previous section, the updates of two probabilities that learn the same evidence converge to the same limiting distribution.

Theorem 1.4. *Let P and Q be probability measures on the same measurable space. Suppose that both sequences $\{P_n : n \in \mathbb{N}\}$ and $\{Q_n : n \in \mathbb{N}\}$ of updates satisfy (M), are uniformly absolutely continuous with respect to their priors, and that P and Q learn the same evidence. Then, $P_\infty = Q_\infty$, where P_∞ and Q_∞ are the respective deterministic limits of $\{P_n : n \in \mathbb{N}\}$ and $\{Q_n : n \in \mathbb{N}\}$.*

Deterministic convergence to the same probability measure is one notion of consensus in the limit. But Theorem 1.4 is not quite a deterministic analogue of the merging of opinions theorem (Section 1.2). We have not shown that $d(P_n, Q_n) \to 0$. Indeed, in the current framework, merging in total variation need not occur.[30] Under an additional, and very mild, assumption about

[30]On the other hand, unlike the Blackwell-Dubins merging of opinions theorem, we do not need to assume absolute continuity for the present result. In many cases, in order to ensure that P and Q are able to learn the same evidence, it may be natural to require P and Q to be mutually absolutely continuous. For if P and Q were not mutually absolutely continuous, it could be the case, for example, that $P(A) = 1$, $Q(A) = 0$, and $A \in \mathcal{F}_1$. Since $P_1 \ll P$ and $Q_1 \ll Q$ by definition, we would have $P_1(A) = 1 \neq 0 = Q_1(A)$, which means that P and Q would not be able to learn the same evidence. Nonetheless, mutual absolute continuity is not needed to prove Theorem 1.4.

probabilistic updates, however, deterministic merging in total variation can be demonstrated.

The mathematical expression of the additional assumption that we need is somewhat technical and has been relegated to the Appendix. But the idea behind the assumption is very simple and intuitive. Until now, an update of P has been *any* probability measure that is absolutely continuous with respect to P. The idea behind the new assumption is that, at each stage n, an update P_n of P should depend on the prior P, the information represented by $\mathcal{F}_n$, *and nothing else*. When this is the case, we will say that P_n is *determined by* P and $\mathcal{F}_n$, and for sequences of updates we will say that $\{P_n : n \in \mathbb{N}\}$ is *determined by* P and $\{\mathcal{F}_n : n \in \mathbb{N}\}$.

Conditionalizations are determined in the sense that we intend. If an agent conditionalizes at stage n, then her posterior probability depends on her prior, the event in $\mathcal{F}_n$ that she learns, and nothing else. I explain this in more detail in the Appendix. The point to emphasize here is that this additional assumption is a very natural one—so natural that it is perhaps surprising that a result like Theorem 1.2 should hold without it. By restricting attention to updates that are determined by P and $\mathcal{F}_n$, we are simply ruling out updates that depend on information beyond that which is encoded in the prior and the available evidence.

With this minor addition, we have the following result.

Theorem 1.5. *Let P and Q be probability measures on the same measurable space. Let $\{P_n : n \in \mathbb{N}\}$ be a sequence of updates of P that is determined by P and $\{\mathcal{F}_n : n \in \mathbb{N}\}$, and let $\{Q_n : n \in \mathbb{N}\}$ be a sequence of updates of Q that is determined by Q and $\{\mathcal{F}_n : n \in \mathbb{N}\}$. Suppose that the sequences $\{P_n : n \in \mathbb{N}\}$ and $\{Q_n : n \in \mathbb{N}\}$ of updates satisfy (M), are uniformly absolutely continuous with respect to their priors, and that P and Q learn the same evidence. Then, $\lim_{n\to\infty} d(P_n, Q_n) = 0$.*

We have demonstrated two senses in which disagreement vanishes as evidence accumulates. Given that the framework of deterministic convergence is able to capture notions of consensus in the limit, I submit that the framework is readily applicable to the issues that were used to motivate the paper.

1.5 Strong Regularity

In Examples 1 and 2 of Section 1.3 a sequence of conditionalizations fails to be uniformly absolute continuous with respect to its prior. We remarked that the key feature of that example is that the

conditionalizations are on events of arbitrarily small prior probability. We are now in a position to say something much more illuminating. When updating goes by conditionalization the assertion that updated probabilities are uniformly absolutely continuous with respect to their prior is *equivalent* to the assertion that the events being conditioned on do not have arbitrarily small prior probability. The latter assertion is represented mathematically, in the following result, using an *infimum*, or greatest lower bound.

Theorem 1.6. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $\{P_n : n \in \mathbb{N}\}$ be a sequence of conditionalizations of P on the events $\{E_n : n \in \mathbb{N}\}$. That is, $P_n = P(\cdot \mid E_n)$ for all $n \in \mathbb{N}$. Then, the sequence $\{P_n : n \in \mathbb{N}\}$ is uniformly absolutely continuous with respect to P if and only if $\inf\{P(E_n) : n \in \mathbb{N}\} > 0$.*

When $\inf\{P(E_n) : n \in \mathbb{N}\} > 0$ holds, we say that the prior probabilities of the events E_n are *bounded away from zero*. To reiterate, this means that the probabilities in question are non-zero *and* do not even approach zero: there is some positive real number that is strictly less than all of them.

In the previous section we saw (Corollary 1.1) that, for updates by conditionalization, uniform absolute continuity is equivalent to deterministic convergence. Combining this result with Theorem 1.6 above, we find that conditionalizations converge deterministically if and only if the prior probabilities of conditioned events are bounded away from zero. We record these equivalences in the following corollary.

Corollary 1.2. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $\{P_n : n \in \mathbb{N}\}$ be a sequence of conditionalizations of P on the events $\{E_n : n \in \mathbb{N}\}$. The following assertions are equivalent.*

(a) The sequence $\{P_n : n \in \mathbb{N}\}$ converges deterministically to a probability measure P_∞ on $(\Omega, \mathcal{F})$.

(b) The sequence $\{P_n : n \in \mathbb{N}\}$ is uniformly absolutely continuous with respect to P.

(c) The probabilities $P(E_n)$, $n \in \mathbb{N}$, are bounded away from zero.

In the remainder of this section, we will show that the above results have interesting connections with other problems in the philosophy of probability.

Let us call a probability *regular* if the only event that it assigns probability zero is the empty—or impossible—event $\emptyset$. Equivalently, a probability is regular if it assigns positive probability to every

non-empty event. There is a well known thesis in the philosophy of probability called *regularity*, which is the view that rationality demands that probabilities be regular (Shimony, 1955; Lewis, 1980; Skyrms, 1980; Hájek, 2011; Easwaran, 2014). A prominent argument for regularity, due to Lewis (1980), is based on the thought that every non-empty proposition in $\mathcal{F}$ is something that can be learned and therefore something on which a Bayesian agent ought to be able to conditionalize. But, according to the ratio definition of conditional probability, this leads to undefined posteriors for irregular probabilities when the proposition learned has prior probability 0. I adopt Easwaran's (2014) statement of the argument.

(P1) Any non-empty proposition in $\mathcal{F}$ can be learned.

(P2) When a rational agent learns E, she conditionalizes on E. That is, she replaces her prior probability P with the the posterior probability $P_E = P(\cdot \mid E)$.

(P3) The conditional probability $P(\cdot \mid E)$ is (by definition) the ratio $P(\cdot \cap E)/P(E)$, and hence undefined when $P(E) = 0$ (see Section 1.2).

(P4) Rational learning cannot leave the posterior probability P_E undefined.

(C1) Therefore, probabilities should be regular.

What is interesting about this argument, in my opinion, is that it derives a synchronic constraint on rational probabilities from a diachronic constraint. From the premise that rational learning goes by conditionalization, we are led to the conclusion that probabilities should be regular. (I return to this point in the final section.) But before we accept the argument's conclusion we should ask about the strength of the constraint that regularity imposes. As Example 1 demonstrates, it is not difficult to construct regular probability measures on countable probability spaces. We need only choose a (suitably normalized) convergent series. But consider the case in which Ω is uncountable, and, for simplicity, suppose that each singleton event $\{\omega\}$ is a member of $\mathcal{F}$. It is a simple mathematical fact that, for any probability P, the set of singletons $\{\omega\}$ with positive probability is countable, and hence P must assign zero probability to uncountably many $\{\omega\}$.[31] If probabilities are to satisfy

[31]The set of singleton events $\{\omega\}$ with positive probability is identical to $\bigcup_{n\in\mathbb{N}}\{\{\omega\} : P(\{\omega\}) > 1/n\}$. And for each $n \in \mathbb{N}$, the set $\{\{\omega\} : P(\{\omega\}) > 1/n\}$ has fewer than n members, else by additivity some finite union of members of this set would exceed 1. Hence, the set of singletons with positive probability is countable because it is a countable union of sets with finite cardinality.

regularity, then they must be defined on countable spaces.

Perhaps probability theory—or, less ambitiously, applications of probability to philosophical problems—can satisfy regularity by eschewing uncountable spaces. But this is extremely implausible. Irregular probability distributions are ubiquitous in mathematical probability theory, statistics, and the sciences. A typical example is the Lebesgue (or uniform) measure on the (closed) unit interval $[0, 1]$. This is the probability measure that assigns to every subinterval of $[0, 1]$ its length. For every real number x in $[0, 1]$, the singleton set $\{x\}$ has Lebesgue measure 0. Moreover, by countable additivity, every countable subset of the unit interval has Lebesgue measure 0. There are also uncountable subsets of the unit interval with Lebesgue measure zero, for example, the Cantor set. Similar observations hold for any probability distribution that is absolutely continuous with respect to Lebesgue measure. Many commonly used probability distributions have this property (e.g. the normal distribution). These examples make it clear that regularity compels its adherents to forsake a substantial portion of probability theory.

So Lewis's argument raises a serious problem. There are several responses to the argument in the literature. For instance, both Lewis (1980) and Skyrms (1980) recommend relaxing the assumption that probability measures are real-valued, allowing probabilities to take values in a hyperreal field containing nonzero infinitesimals. It is possible to assign uncountably many singletons non-zero, infinitesimal probability, thereby satisfying regularity. This recommendation is discussed at length by Easwaran (2014).

Another way to respond to the argument is to deny (P3) and use a theory of conditional probability that permits conditioning on probability 0 events. There are several such theories in the literature (Popper, 1955; Renyi, 1970; Dubins, 1975). According to these theories, conditional probabilities are not *defined* as ratios of unconditional probabilities. Rather, conditional probabilities are treated as primitives that *satisfy* the ratio condition when the conditioning event has positive probability.[32]

Finally, another fairly mainstream response to Lewis's argument for regularity is to reject the view that rational learning always goes by conditionalization. Several alternatives to conditionalization are mentioned in footnote 22.

[32]See Seidenfeld (2001) for further discussion.

The results of this paper suggest a new argument that is similar to Lewis's but with a stronger conclusion. Let us call a probability *strongly regular* if, for any decreasing sequence of non-empty events, the probabilities of the events in the sequence are bounded away from zero. As the terminology suggests, if a probability is strongly regular, then it is regular.[33] But, as we have already seen, there are probabilities that are regular and not strongly regular. In Example 1, the probability P assigns non-zero probability to every non-empty event, but there is a decreasing sequence of events whose probabilities are not bounded away from zero.[34] In analogy with regularity, let *strong regularity* be the thesis that rationality demands that probabilities be strongly regular. We can now write an argument for strong regularity.

(P5) Any decreasing sequence of non-empty events in $\mathcal{F}$ can be learned.

(P2) When a rational agent learns E, she conditionalizes on E.

(P3) Conditional probabilities are defined using the standard ratio definition.

(P6) If a rational agent learns increasing and complete evidence and (M) holds, then her updated probabilities converge deterministically.

(C2) Therefore (by Corollary 1.2), rational probabilities are strongly regular.

Here is another way of putting the argument. Suppose for contradiction that there is a rational probability measure that is not strongly regular. As any sequence of events may be learned (P5), let the agent learn a sequence whose prior probabilities tend to zero. Since, by (P2), learning is by conditionalization, (M) is satisfied. By (P6), the agent's updated probabilities converge deterministically. But this contradicts the lack of strong regularity, by Corollary 1.2. Therefore, rational probabilities are strongly regular.

Notice that the conclusion (C2) gives rise to an even more severe cardinality bound on probability spaces than the conclusion (C1). Not only does strong regularity rule out uncountable spaces (since it implies regularity), it also rules out *countably infinite* spaces in the following fashion.

[33] Consider the constant sequences $\{A : n \in \mathbb{N}\}$ for each non-empty $A \in \mathcal{F}$.

[34] An anonymous referee has suggested the following, helpful characterization of strongly regular probabilities: If a probability is regular, then it is strongly regular if and only if the intersection of any decreasing sequence of non-empty events is non-empty. This follows from the continuity of countably additive probabilities. Note that the probability measure in Example 1 violates this condition.

Suppose Ω is countable and that, as above, each singleton $\{\omega\}$ is a member of $\mathcal{F}$. If the probability space $(\Omega, \mathcal{F}, P)$ is strongly regular, then for some natural number n, all singletons satisfy $P(\{\omega\}) \geq 1/n$. If this were not the case, then there would exist (as in Example 1) a decreasing sequence of events with probabilities not bounded away from zero.[35] But this implies that there are at most n singleton events because the sum of the singletons' probabilities cannot exceed 1. So our probability space must be finite (with cardinality at most n).

Note also that, unlike the argument for regularity, relaxing the requirement that probabilities be real-valued does not answer the argument for strong regularity. Strong regularity requires that probabilities be strictly greater than some positive *real number*. But there are no infinitesimals with this property. So, even if Bayesians allow probabilities to take infinitesimal values, it is still not possible to define strongly regular probabilities on countably infinite and uncountable spaces.

It seems to me that the most controversial premise of the argument is (P6). But (P6) is not, I think, implausible because it only asks for deterministic convergence in the most ideal learning scenarios.[36] Given an increasing stream of evidence that eventually informs one about every event that one is interested in and a procedure for updating probabilities that never contradicts previous updates (condition (M)), it is not unreasonable to hope that rational learning leads one's probability judgments to approach *some* limit in the long run. The hope is simply that, in these highly idealized learning scenarios, one's update procedures do not produce probabilities that oscillate indefinitely. But as we have seen, it is exceedingly difficult for Bayesian learning to realize this small hope, and, therefore, exceedingly difficult for Bayesians to use deterministic convergence to underwrite notions of objectivity and rational consensus.

How might Bayesians respond to this argument? As was the case with Lewis's argument for regularity, one response is to reject the view that rational learning always goes by conditionalization. I leave it as an open question whether alternative update procedures require strong regularity or similar properties. As was also the case with Lewis's argument, one may wish to reject (P3) and adopt an alternative theory of conditional probability. It is essential to the proof of Theorem 1.6 that conditional probabilities be defined as ratios of unconditional probabilities. I do not currently know

[35] Formally, with the sequence of events $\{E_n : n \in \mathbb{N}\}$ defined as in Example 1, this follows from the following elementary fact about the tails of a non-negative convergent series: if $\sum_{i=1}^{\infty} P(\{\omega_i\}) < \infty$, then $P(E_n) = \sum_{i=n+1}^{\infty} P(\{\omega_i\}) \to 0$ as $n \to \infty$.

[36] For a recent endorsement of something like (P6), see Autzen (2018).

the consequences of retaining (P2), (P5), and (P6) and using primitive conditional probabilities.

A more radical response is to reject (P6) by way of altogether denying that a satisfying epistemology needs to provide accounts of scientific objectivity and intersubjective agreement. The idea here is simply to bite the bullet in response to the charge that Bayesianism is too subjective. This line of response seems to be endorsed by some theorists. For example, Joseph Kadane has said that claims of objectivity are "insupportable" and that "statements about the probabilities of specific events are representations of the opinions of the writer, i.e. they are personal" (Kadane, 2009, p. 110). Less radical Bayesians who are not willing to embrace complete subjectivism face a difficult problem. If asymptotic results are to underwrite accounts of objectivity and intersubjective agreement, then either Bayesians must settle for the almost sure qualifications that Glymour, Earman, Belot, and others have attacked, or they must embrace the consequences of imposing deterministic convergence, including strong regularity.

1.6 Conclusion

To summarize, this paper's contributions have been, first, to raise a natural question about Bayesian convergence results that has received relatively little attention in the philosophical literature, namely: Under what conditions do probabilistic updates converge deterministically? Second, we have provided a simple, but fairly conclusive, answer to that question at a high level of generality. Finally, we have shown that this answer has interesting connections to other problems in the philosophy of probability. If Bayesian learning is to produce deterministic convergence, then agents' priors must satisfy the extremely demanding strong regularity thesis.

Although the main points have been somewhat negative in character, pointing out difficulties associated with deterministic convergence, I will conclude on a more positive note. Bayesian theory is sometimes described as consisting of a synchronic component and a diachronic component. The synchronic component is the thesis that rational degrees of belief are probabilistically coherent, and the diachronic component is the thesis that rational learning goes by conditionalization on evidence. One lesson of this paper is that these two components can exhibit significant interdependence. In particular, we have seen that one's theory of learning can place significant constraints on one's theory of synchronic rationality: if learning is to bring about deterministic convergence, then Bayesian

probabilities should be strongly regular. This interdependence suggests several questions for future research. For example, what kinds of synchronic constraints arise from well-known generalizations of Bayesian conditionalization, like probability kinematics? Or, what kinds of learning procedures are consistent with irregular probability assignments? And what asymptotic properties do they have?

Appendix

The Vitali-Hahn-Saks Theorem

The Vitali-Hahn-Saks Theorem is a general measure-theoretic result that does not depend on the measures in question being probabilities. A *finite measure space* $(\Omega, \mathcal{F}, \mu)$ has the same properties as a probability space, except we no longer require $\mu(\Omega) = 1$, but only that $\mu(\Omega) < \infty$.

Theorem 1.7 (Vitali-Hahn-Saks)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a finite measure space and $\{\mu_n : n \in \mathbb{N}\}$ a sequence of finite measures on $(\Omega, \mathcal{F})$ such that $\mu_n \ll \mu$ for all $n \in \mathbb{N}$. Suppose that the sequence $\{\mu_n(\Omega) : n \in \mathbb{N}\}$ is bounded and that $\{\mu_n : n \in \mathbb{N}\}$ converges deterministically to the set function μ_∞. Then the sequence $\{\mu_n : n \in \mathbb{N}\}$ is uniformly absolutely continuous with respect to μ. Moreover, μ_∞ is a finite measure on $(\Omega, \mathcal{F})$ with $\mu_\infty \ll \mu$.*

Proof. See Royden and Fitzpatrick (2010, Section 18.5). □

Proof of Theorem 1.2

Proof. The proof of this theorem relies heavily on martingale theory. We appeal to several standard results without giving their proofs. References will be provided instead.

Let $P|_{\mathcal{F}_n}$ ($P_n|_{\mathcal{F}_n}$) be the restriction of P (P_n) to the measurable space $(\Omega, \mathcal{F}_n)$. Since $P_n \ll P$ for all n, we may define $Z_n = dP_n|_{\mathcal{F}_n}/dP|_{\mathcal{F}_n}$ to be the corresponding Radon-Nikodym derivatives on $(\Omega, \mathcal{F}_n)$ so that Z_n is $\mathcal{F}_n$-measurable. Then, for $F \in \mathcal{F}_{n-1}$, we have

$$\int_F Z_{n-1} dP = P_{n-1}(F).$$

But if $F \in \mathcal{F}_{n-1}$, then $F \in \mathcal{F}_n$ as well, hence

$$\int_F Z_n dP = P_n(F).$$

By condition (M), $P_{n-1}(F) = P_n(F)$, so

$$\int_F Z_{n-1} dP = \int_F Z_n dP.$$

This last equation shows that the sequence $\{Z_n : n \in \mathbb{N}\}$ is a non-negative martingale in $\{\mathcal{F}_n : n \in \mathbb{N}\}$ (Durrett, 2010, Section 4.2). Therefore, by the martingale convergence theorem (Durrett, 2010, Section 4.2, Theorem 2.10), $\{Z_n : n \in \mathbb{N}\}$ converges almost surely to an integrable random variable Z_∞ as $n \to \infty$.

Moreover, since $\{P_n : n \in \mathbb{N}\}$ is uniformly absolutely continuous with respect to P, the restricted sequence $\{P_n|_{\mathcal{F}_n} : n \in \mathbb{N}\}$ is uniformly absolutely continuous with respect to P in the sense that for all $\epsilon > 0$ there exists a $\delta > 0$ such that for all n and all $F \in \mathcal{F}_n$,

$$P_n(F) < \epsilon \quad \text{whenever} \quad P(F) < \delta.$$

This slight variant of the textbook definition of uniform absolute continuity, given in the main text, is studied by Huttegger (2015b), who shows (Lemma 12.1) that uniform absolute continuity in this sense implies that $\{Z_n : n \in \mathbb{N}\}$ is uniformly integrable with respect to P (this result holds for the textbook definition too (Royden and Fitzpatrick, 2010, Section 18.5, Proposition 24)).

Now, for $F \in \bigcup_n \mathcal{F}_n$ and all sufficiently large n we have

$$P_n(F) = \int_F Z_n dP,$$

and, by the Vitali convergence theorem (Royden and Fitzpatrick, 2010, Section 18.3),

$$\lim_{n\to\infty} P_n(F) = \lim_{n\to\infty} \int_F Z_n dP = \int_F Z_\infty dP.$$

As Z_∞ is a non-negative P-integrable function, P_∞ defined by $P_\infty(A) = \int_A Z_\infty dP$ is a probability measure on $\mathcal{F}$. Hence, $\{P_n(F) : n \in \mathbb{N}\}$ converges to $P_\infty(F)$ for all $F \in \bigcup_n \mathcal{F}_n$. It remains to show that $\{P_n(A) : n \in \mathbb{N}\}$ converges to $P_\infty(A)$ for all $A \in \mathcal{F}$.[37]

Since, $\bigcup_n \mathcal{F}_n$ is a π-system that generates $\mathcal{F}$, it suffices to show that

$$\mathcal{C} = \{A \in \mathcal{F} : \lim_{n\to\infty} P_n(A) = P_\infty(A)\}$$

[37]Thanks to saz at math.stackexchange for helping me to verify this step of the proof. Any mistakes, of course, are mine.

contains Ω and is closed under complementation and disjoint countable unions, for then our result follows by Dynkin's π-λ theorem. Clearly, $\Omega \in \mathcal{C}$ because for all n, $P_n(\Omega) = P_\infty(\Omega) = 1$. If $A \in \mathcal{C}$, then

$$\lim_{n\to\infty} P_n(A^c) = 1 - \lim_{n\to\infty} P_n(A) = 1 - P_\infty(A) = P_\infty(A^c),$$

so $\mathcal{C}$ is closed under complementation. Now let $\{A_k : k \in \mathbb{N}\}$ be a sequence of pairwise disjoint events in $\mathcal{C}$ and write $A = \bigcup_k A_k$. Let $\epsilon > 0$ be arbitrary and use the uniform absolute continuity of $\{P_n : n \in \mathbb{N}\}$ with respect to P to find a $\delta > 0$ such that for all n and all $A \in \mathcal{F}$,

$$P(A) < \delta \text{ implies } P_n(A) < \epsilon/4. \tag{1.4}$$

Using the fact that P and P_∞ are countably additive, and so continuous, there exists $K \in \mathbb{N}$ such that

$$P\big(A - \bigcup_{k=1}^{K} A_k\big) \le \delta \text{ and } P_\infty\big(A - \bigcup_{k=1}^{K} A_k\big) \le \epsilon/4.$$

Then, using the triangle inequality and (1.4), we have

$$\begin{aligned}
|P_n(A) - P_\infty(A)| &= \Big|P_n\big(A - \bigcup_{k=1}^{K} A_k\big) + P_n\big(\bigcup_{k=1}^{K} A_k\big) - P_\infty\big(A - \bigcup_{k=1}^{K} A_k\big) - P_\infty\big(\bigcup_{k=1}^{K} A_k\big)\Big| \\
&\le \Big|P_n\big(A - \bigcup_{k=1}^{K} A_k\big) - P_\infty\big(A - \bigcup_{k=1}^{K} A_k\big)\Big| + \Big|P_n\big(\bigcup_{k=1}^{K} A_k\big) - P_\infty\big(\bigcup_{k=1}^{K} A_k\big)\Big| \\
&\le \Big|P_n\big(A - \bigcup_{k=1}^{K} A_k\big) - P_\infty\big(A - \bigcup_{k=1}^{K} A_k\big)\Big| + \sum_{k=1}^{K} |P_n(A_k) - P_\infty(A_k)| \\
&\le \epsilon/2 + \sum_{k=1}^{K} |P_n(A_k) - P_\infty(A_k)|.
\end{aligned}$$

Since $A_k \in \mathcal{C}$, we have

$$\lim_{n\to\infty} \sum_{k=1}^{K} |P_n(A_k) - P_\infty(A_k)| = 0,$$

and therefore, for all but finitely many n,

$$|P_n(A) - P_\infty(A)| \le \epsilon.$$

This shows that $A \in \mathcal{C}$ because ϵ is arbitrary. We can now conclude that $\{P_n(A) : n \in \mathbb{N}\}$ converges

to $P_\infty(A)$ for all $A \in \mathcal{F}$. □

Proof of Theorem 1.4

Proof. Given the assumptions, we know from Theorem 1.2 that the deterministic limits P_∞ and Q_∞ of the sequences of updates $\{P_n : n \in \mathbb{N}\}$ and $\{Q_n : n \in \mathbb{N}\}$, respectively, exist. We also know that P_∞ and Q_∞ are probability measures on $(\Omega, \mathcal{F})$. We show that P_∞ and Q_∞ assign the same probability to events in the algebra $\bigcup_n \mathcal{F}_n$ that generates $\mathcal{F}$. From this it follows, by a standard generating class argument, that $P_\infty = Q_\infty$.

Let $A \in \bigcup_n \mathcal{F}_n$, and assume $A \in \mathcal{F}_{n_0}$. Since P and Q learn the same evidence,

$$P_{n_0}(A) = Q_{n_0}(A),$$

and by (M),

$$P_n(A) = P_{n_0}(A) \text{ and } Q_n(A) = Q_{n_0}(A)$$

for all $n \geq n_0$. Hence, $P_n(A) = Q_n(A)$ if $n \geq n_0$, which implies $P_\infty(A) = Q_\infty(A)$ for arbitrary $A \in \bigcup_n \mathcal{F}_n$. This establishes the desired result. □

Proof of Theorem 1.5

The notion of determination from the main text is expressed mathematically as follows. If P_n is an update of P, then by assumption $P_n \ll P$. Therefore, the Radon-Nikodym derivative $Z_n := dP_n/dP$ exists. We say that P_n *is determined by* P and $\mathcal{F}_n$ if Z_n is $\mathcal{F}_n$-measurable, i.e. for all Borel subsets B of $\mathbb{R}$, $\{\omega \in \Omega : Z_n(\omega) \in B\} \in \mathcal{F}_n$.

In the main text, it is assumed that $\mathcal{F}_n$ is generated by a finite partition $\mathcal{E} = \{E_1, ..., E_k\}$. We claimed that conditionalizations of P are determined by P and $\mathcal{F}_n$. To verify this, suppose that $P_n = P(\cdot \mid E_j)$, $E_j \in \mathcal{E}$. Then, $Z_n = \mathbf{1}_{E_j}/P(E_j)$. Hence, Z_n is constant on the atoms of $\mathcal{F}_n$. That is, $Z_n(\omega) = 1/P(E_j)$ if $\omega \in E_j$ and $Z_n(\omega) = 0$ if $\omega \in E_i$, $i \neq j$. This implies that Z_n is $\mathcal{F}_n$-measurable.

The proof of Theorem 1.5 goes by way of a strengthening of Theorem 1.2. To prove this result, we will appeal to a well-known fact about total variation distance. Let P and Q be probability

measures on $(\Omega, \mathcal{F})$, and let m be any measure such that $P \ll m$ and $Q \ll m$. Such an m always exists, for example, $m = P/2 + Q/2$. Let $p = dP/dm$ and $q = dQ/dm$ be the Radon-Nikodym derivatives of P and Q with respect to m. Then,

$$d(P, Q) = \frac{1}{2} \int |p - q| dm. \tag{1.5}$$

Theorem 1.8. *Let $(\Omega, \mathcal{F}, P)$ be a probability space equipped with a filtration $\{\mathcal{F}_n : n \in \mathbb{N}\}$ that generates $\mathcal{F}$. Let $\{P_n : n \in \mathbb{N}\}$ be a sequence of updates of P that satisfies (M) and is determined by P and $\{\mathcal{F}_n : n \in \mathbb{N}\}$. If $\{P_n : n \in \mathbb{N}\}$ is uniformly absolutely continuous with respect to P, then P_n* uniformly *converges deterministically to a probability measure P_∞ on $(\Omega, \mathcal{F})$. That is, $\lim_{n\to\infty} d(P_n, P_\infty) = 0$.*

Proof. We begin by noting that, under the assumption that $\{P_n : n \in \mathbb{N}\}$ is determined by P and $\{\mathcal{F}_n : n \in \mathbb{N}\}$, the proof of Theorem 1.2 goes through verbatim by setting $Z_n := dP_n/dP$ for all $n \in \mathbb{N}$. In particular, the sequence $\{Z_n : n \in \mathbb{N}\}$ is a non-negative, uniformly integrable martingale in $\{\mathcal{F}_n : n \in \mathbb{N}\}$. Therefore, there exists an integrable Z_∞ random variable on $(\Omega, \mathcal{F})$ such that $Z_n \to Z_\infty$ almost surely, and with P_∞ defined by $P_\infty(A) = \int_A Z_\infty dP$ for all $A \in \mathcal{F}$, it follows from the Vitali convergence theorem that $\{P_n : n \in \mathbb{N}\}$ converges deterministically to P_∞ and $P_\infty \ll P$.

By (1.5), it suffices to show that

$$\lim_{n\to\infty} \int |Z_n - Z_\infty| dP = 0. \tag{1.6}$$

But it is well-known that the martingale $\{Z_n : n \in \mathbb{N}\}$ is uniformly integrable if and only if (1.6) holds (Durrett, 2010, Section 4.5, Theorem 5.6). So we are done. □

Finally, we have

Proof of Theorem 1.5. By Theorem 1.8, there exist probabilities P_∞ and Q_∞ such that

$$\lim_{n\to\infty} d(P_n, P_\infty) = 0 \text{ and } \lim_{n\to\infty} d(Q_n, Q_\infty) = 0.$$

By Theorem 1.4, $P_\infty = Q_\infty$, hence $d(P_\infty, Q_\infty) = 0$. By the triangle inequality,

$$d(P_n, Q_n) \leq d(P_n, P_\infty) + d(P_\infty, Q_\infty) + d(Q_n, Q_\infty) \to 0 \text{ as } n \to \infty,$$

and the proof is complete. □

Proof of Theorem 1.6

Proof. First, suppose it is not the case that $\alpha := \inf\{P(E_n) : n \in \mathbb{N}\} > 0$. Then, since in general $\alpha \geq 0$, we have $\alpha = 0$. We want to show that $\{P_n : n \in \mathbb{N}\}$ is not uniformly absolutely continuous with respect to P, that is, that there exists an $\epsilon > 0$ such that for all $\delta > 0$ there exists an $n \in \mathbb{N}$ and an event $A \in \mathcal{F}$ such that $P(A) < \delta$ and $P_n(A) \geq \epsilon$. Let $\epsilon = 1/2$ and let $\delta > 0$ be given. As $\alpha = 0$, there exists an $n \in \mathbb{N}$ such that $P(E_n) < \delta$. But, for such n, $P_n(E_n) = 1 \geq 1/2$ because P_n comes from conditionalizing P on E_n.

Conversely, suppose $\alpha > 0$ and let $\epsilon > 0$ be given. Let $\delta := \epsilon\alpha > 0$. As α is the infimum over $P(E_n)$, we have $\delta \leq \epsilon P(E_n)$ for all $n \in \mathbb{N}$. So, for all $n \in \mathbb{N}$ and $A \in \mathcal{F}$, $P(A) < \delta$ implies $P(A \cap E_n) < \delta$, which in turn implies

$$P_n(A) = \frac{P(A \cap E_n)}{P(E_n)} < \frac{\delta}{P(E_n)} \leq \frac{\epsilon P(E_n)}{P(E_n)} = \epsilon,$$

as desired. □

Chapter 2

Persistent Disagreement and Polarization in a Bayesian Setting

2.1 Introduction

In politics, group deliberation, and interpersonal relationships, persistent disagreement and belief polarization are often seen as lamentable features of social life. In some cases, polarization can even be dangerous. "Terrorism itself is a product, in part, of group polarization," as Cass Sunstein points out (2002, p. 187). In certain politicized areas of science—such as vaccination, minimum wage policy, and climate change—disagreement is commonly attributed to the irrationality of one party to the debate.[1] If only we were rational, one might think, all disagreement would be resolved by collecting and sharing evidence. Call this *the optimistic thesis about learning* (TOTAL).

TOTAL. Rational agents who learn the same evidence resolve disagreements.

Something like TOTAL, made suitably precise, seems to underwrite many of our practices in diverse areas like activism, argumentation, inquiry, and mediation.[2] It might be quite hard to explain our actions in those areas in a non-cynical fashion without TOTAL.[3] The thesis also motivates positions in popular philosophical disputes. Conciliatory positions in the peer disagreement debate take it that "the peer's disagreement gives one evidence that one has made a mistake in interpreting the original evidence, and that such evidence should diminish one's confidence" on the issue about which there is disagreement (Christensen, 2009, p. 757). If TOTAL were false, why would mere

[1] Not always. Sometimes disagreement is attributed to a profit motive, for example. But good faith and rationality attributions to opponents in such debates are fairly rare.

[2] An anonymous referee has suggested that proponents of TOTAL be called *totalitarians*.

[3] So-called "virtue signaling" is an example of a cynical explanation in the case of activism.

disagreement diminish one's confidence? Two agents could disagree in the face of shared evidence without either having made a mistake.

But TOTAL *is* false. In fact, even certain weaker theses are untenable. For example, versions of TOTAL that require shared evidence to guard against polarization—the case in which the extent of disagreement increases—cannot be maintained. Here we study persistent disagreement and belief polarization in a general Bayesian framework. Many formal studies of social opinion dynamics focus on reaching consensus (e.g., Blackwell and Dubins, 1962; DeGroot, 1974; Lehrer and Wagner, 1981; Genest and Zidek, 1986). But persistent disagreement and polarization are interesting phenomena worthy of study in their own right. While modeling these phenomena has been addressed in the literature (e.g., Hegselmann et al., 2002), we want to study them in a Bayesian setting. The point of adopting this framework is that Bayesian learning represents a leading contender for an *ideal* standard of rational belief revision. Even in this setting, both persistent disagreement in general and belief polarization in particular are possible in the face of increasing, *shared* evidence.

We make our case against TOTAL in two parts. In Part I, we study persistent disagreement and polarization in circumstances in which agents learn some finite amount of shared evidence. This is an important case since it resembles the contexts in which we observe actual belief polarization. Our primary focus in Part 2.1 is polarization, of which we distinguish two senses. Since polarization (in either of our senses) implies persistent disagreement, cases in which polarization is rationally permissible represent clear failures of TOTAL. The first sense of polarization that we investigate (Section 2.2) is local in that it involves increases in the extent of disagreement *with respect to a particular event of interest*. We provide some simple characterizations of precisely when polarization in this sense occurs (Theorem 2.1). From a Bayesian perspective, no irrationality is required. In Section 2.3, we discuss a close connection between polarization with respect to an event and dilation, a well-studied phenomenon in the theory of imprecise probabilities. In Section 2.4 we introduce a global notion of polarization that does not depend on a particular event. When agents polarize globally, the *overall* extent to which their probability distributions disagree increases. As with polarization with respect to an event, global polarization is sometimes permissible for Bayesian agents.

In Part II, we argue against an even weaker version of TOTAL by turning to the general, idealized setting of Bayesian learning in which agents are able to learn an infinite amount of evidence.

Our main mathematical contribution there is a result that we call the *Bayesian Consensus-or-Polarization Law*. That result generalizes the classic merging of opinions result due to Blackwell and Dubins (1962) by relaxing a heavy-handed assumption (absolute continuity) in a mild way with interesting consequences. Relative to our weaker assumption, while it is no longer the case that an agent must assign probability 1 to achieving consensus in the limit, she must assign probability 1 to either achieving consensus or polarizing. So not only is polarization consistent with rationality in the sorts of learning scenarios in which it is observed (Part 2.1), in some cases, *rationality demands assigning positive probability* to *polarizing* in the limit of inquiry (Part 2.4).[4]

Part I: The Finite Case

Both psychological evidence and common life experience attest to the reality of persistent disagreement and belief polarization. In a classic study on polarization, Lord et al. report that, when exposed to the same set of conflicting studies regarding the possible deterrent effects of the death penalty, subjects disagreeing initially strengthened their respective views, coming to disagree even more strongly (1979). Ross and Anderson claim that this behavior stands "in contrast to any normative strategy imaginable for incorporating new evidence relevant to one's belief" (1982, p. 145).

In order to investigate the rational status of such observed behavior, we begin by looking at learning situations that fairly closely approximate those in which the behavior occurs. In particular, we first consider cases of learning *finitely many* events. We focus exclusively on Bayesian conditionalization—according to which *events* are learned with *certainty*—rather than Jeffrey conditioning or more general rules (Jeffrey, 2004). If Bayesian learning cannot guard against persistent disagreement and belief polarization, then generalizations cannot either.[5]

[4]Another idea worth exploring in connection with this paper's arguments is that rationality might demand that an agent assign positive probability to every event of interest to her. In other words, rationality might forbid making positive probability assignments to too many alternative hypotheses. Gaifman and Snir (1982) provide precise results bearing on this idea: the more hypotheses that an agent tries to accommodate, the more complicated her prior becomes.

[5]One might ask, as an anonymous referee did, whether in contexts of *only* uncertain evidence—so that updates by Jeffrey conditioning don't reduce to standard conditionalization—persistent disagreement and polarization can be avoided. In general, the answer is no. One subtle issue in this context is how to understand shared evidence. See Huttegger (2015b) for a study of conditions that secure merging of opinions for Jeffrey conditioning. In particular, see his Theorem 6.1 for a class of cases in which disagreement persists.

2.2 Polarization

We begin with the phenomenon of belief polarization. Since, as we will show, polarization is possible for rational agents who learn the same finite amount of shared evidence, it follows that persistent disagreement is also possible. Polarization represents a radical failure of TOTAL. While it may be familiar from debates about uniqueness and permissivism that standard varieties of Bayesianism allow for persistent disagreement of some form (e.g., White, 2005), it does not follow from that fact that Bayesianism allows for belief polarization.

Throughout the paper, let Ω be a set of *elementary events* or *possible worlds*. Let $\mathcal{F}$ be a sigma-algebra on Ω, that is, a non-empty collection of subsets of Ω closed under complementation and countable unions. Elements $A \in \mathcal{F}$ are called *events*. To take a simple and common example, Ω may contain six points representing the outcomes of a roll of a die, and $\mathcal{F}$ might contain all of the relevant die-rolling events such as $\{2, 4, 6\}$, the event that the die lands with an even number face up. A *probability measure* P on the *measurable space* $(\Omega, \mathcal{F})$ is a countably additive set function $P : \mathcal{F} \to [0, 1]$ such that $P(\Omega) = 1$.[6] We assume the standard ratio definition of conditional probability. For all $A, E \in \mathcal{F}$,

$$P(A|E) := \frac{P(A \cap E)}{P(E)}, \text{ whenever } P(E) > 0.$$

According to Bayesian conditionalization, when an agent learns an event E, she should revise her probabilities by setting her new probabilities equal to her old probabilities conditional on E. Where P^E is the probability measure that the agent adopts after learning E, conditionalization says that

$$P^E(A) = P(A|E),$$

for all $A \in \mathcal{F}$. We call P^E the *posterior* and P the *prior*.

With those few preliminaries out of the way, consider the following simple example.

Example 2.1. *Suppose two polling experts provide opinions about the outcome of an election between four candidates. Let $\Omega = \{\omega_1, \omega_2, \omega_2, \omega_4\}$ be the set of candidates, and let $\mathcal{F} = 2^\Omega$. Let P_1,*

[6] We say that P is *countably additive* if, for any countable collection of disjoint events $\{A_i\}_{i \in I}$, $P(\bigcup_{i \in I} A_i) = \sum_{i=1}^{\infty} P(A_i)$. We note that our Theorem 2.3 makes essential use of countable additivity. For the simple examples considered in this section, however, finite additivity is sufficient.

P_2, given by Table 2.1, be probability measures on $(\Omega, \mathcal{F})$ representing the opinions of the polling experts.

Table 2.1: Priors

	ω_1	ω_2	ω_3	ω_4
P_1	1/6	1/4	1/3	1/4
P_2	1/2	1/12	1/4	1/6

Suppose that there are two political parties with ω_1, ω_2 in one party, and ω_3, ω_4 in the other. In order to advance to the general election, a candidate must win her party's primary. Suppose that we are interested in the event $A = \{\omega_1, \omega_2\}$ that a candidate from the first party will win the general election. Let $E = \{\omega_1, \omega_3\}$ be the event that candidates 1 and 3 win their respective primaries. Then, $P_1(\cdot|E)$ and $P_2(\cdot|E)$ are given by Table 2.2.

Table 2.2: Posteriors

	ω_1	ω_2	ω_3	ω_4
P_1^E	1/3	0	2/3	0
P_2^E	2/3	0	1/3	0

Then, $P_1(A|E) = 1/3 < 5/12 = P_1(A) \leq P_2(A) = 7/12 < 2/3 = P_2(A|E)$. So, updating on E, P_1 and P_2 get further apart with respect to A. That is, the information that candidates 1 and 3 win their primaries pushes the two polling experts further apart with respect to their opinions about whether a candidate from the first party will win the general election. △

The behavior in Example 2.1 is eminently reasonable. Given polling expert 1's prior, the information $E = \{\omega_1, \omega_3\}$ is tantamount to learning that the weaker candidate from the first party will run against the stronger candidate from the second party in the general election. Indeed, expert 1 considers candidate ω_1 antecedently weaker than *all* other candidates. For expert 2,

however, the case is reversed. Not only is ω_1 the stronger candidate in the first party, she is the strongest candidate overall. So while E decreases expert 1's confidence in A, the event that a candidate from the first party wins the general election, it increases expert 2's confidence. Nothing in this example or the next (Example 2.2) depends on unreasonable or extreme (0-1 valued) prior probability assignments.

Example 2.1 suggests the following natural definition of polarization.

Definition 2.1. Let P_1 and P_2 be probability functions on $(\Omega, \mathcal{F})$, and let $A, E \in \mathcal{F}$. We say that evidence E *polarizes* P_1 *and* P_2 *with respect to the event* A if

$$P_1(A|E) < P_1(A) \leq P_2(A) < P_2(A|E).$$

Polarizing evidence leads disagreeing opinions to strengthen their initial attitude with respect to each other, resulting in even greater disagreement. Note that Definition 2.1 is given in terms of a particular event with respect to which polarization occurs. We will later (Section 2.4) distinguish this sense of polarization from a sense that does not depend on a particular event.

Example 2.1 already shows that polarization is rationally permissible from a Bayesian perspective. To shed more light on the situation, we would now like to provide some simple conditions that characterize when polarization with respect to an event occurs. We begin by introducing some auxiliary probabilistic concepts. First, when $P(A \cap E) = P(A)P(E)$, we call events A and E *stochastically independent* (according to P). Next, consider the function S defined by

$$S_P(A, E) := \frac{P(A \cap E)}{P(A)P(E)}.$$

The covariance of A and E is given by $Cov(A, E) = P(A \cap E) - P(A)P(E)$. S just puts covariance in ratio form. When $S_P(A, E) = 1$, A and E are stochastically independent. When $S_P(A, E) > 1$, A and E are positively correlated. And A and E are negatively correlated when $S_P(A, E) < 1$. (Seidenfeld and Wasserman adopt the convention that $S_P(A, E) = 1$ when $P(A)P(E) = 0$ (1993, p. 1141).) Pedersen and Wheeler point out (2014, p. 1312, fn. 8) that S has been put to various uses in formal epistemology and philosophy of science, including as a measure of coherence (Shogenji, 1999). Finally, the quantity $P(E|A)/P(E|A^c)$ is called the *likelihood ratio* for data E

and hypotheses A and A^c. Likelihood ratios are used throughout statistics and express the impact of the evidence in terms of how much it favors one hypothesis to another.

Together, S and the likelihood ratio allow us to state two very simple characterizations of polarization with respect to an event.

Theorem 2.1. *Suppose that* $0 < P_1(A) \leq P_2(A) < 1$. *Then the following are equivalent.*

(i) Evidence E polarizes P_1 and P_2 with respect to A;

(ii) $S_{P_1}(A, E) < 1 < S_{P_2}(A, E)$;

(iii) $\dfrac{P_1(E|A)}{P_1(E|A^c)} < 1 < \dfrac{P_2(E|A)}{P_2(E|A^c)}$.

The proof of this result uses only the probability axioms and algebra. We omit it, assured the reader can furnish it herself should she so desire.

Condition (iii) of Theorem 2.1 is mentioned in psychological literature on polarization (e.g., Jern et al., 2014). Jern et al. analyze previous empirical studies of belief polarization and offer a Bayesian rationalization of some such behavior. They also note that such likelihood ratios as included in condition (iii) determine "the direction in which [an agent's] beliefs will change." Condition (ii) has been exploited in another, related literature that we discuss below (e.g., Seidenfeld and Wasserman, 1993; Wasserman and Seidenfeld, 1994; Pedersen and Wheeler, 2014). Theorem 2.1, while neither particularly deep nor surprising, is more general than it may seem at first. That's because conditionalizing on any finite string of evidence, $E_1, ..., E_n$, can be reduced to learning just a single event, namely, $E = \bigcap_{i=1}^{n} E_i$. So Theorem 2.1 identifies necessary and sufficient conditions for any finite string of evidence to polarize P_1 and P_2 with respect to A.[7] If $(\Omega, \mathcal{F})$ is a sufficiently complex space relative to how much the agents learn—say there is a countable infinity of events, for example, and agents only learn a finite amount of evidence—the agents may be polarized by the event representing their total evidence.

[7]We could distinguish the case in which $\bigcap_{i=1}^{n} E_i$ polarizes P_1 and P_2 with respect A from the case in which each E_i polarizes P_1 and P_2 with respect to A. In the former case, we are concerned with the cumulative effect of the evidence. The latter case obtains when the characterizing conditions of Theorem 2.1 hold for each piece of evidence, E_i.

We offer another, more extreme example of polarization adapted from (Herron et al., 1994; Pedersen and Wheeler, 2015).

Example 2.2. *Consider P_1 and P_2 such that $P_1(G) = 0.1$ and $P_2(G) = 0.9$. Consider the toss of a coin that is fair according to both P_1 and P_2: $P_1(H) = P_2(H) = 1/2 = P_1(H^c) = P_2(H^c)$. Suppose that the outcomes of the coin toss are independent of the event G according to both P_1 and P_2. So, $P_1(G \cap H) = P_1(G)P_1(H)$ and $P_2(G \cap H) = P_2(G)P_2(H)$. Let A be the "matching" event that either both G and H occur or both do not. That is, $A := (G \cap H) \cup (G^c \cap H^c)$. Notice that $P_1(A) = 1/2 = P_2(A)$. Despite initial agreement concerning A, the coin toss polarizes $P_1(A)$ and $P_2(A)$. For $i = 1, 2$,*

$$P_i(A|H) = \frac{P_i([(G \cap H) \cup (G^c \cap H^c)] \cap H)}{P_i(H)} = \frac{P_i(G \cap H)}{P_i(H)} = \frac{P_i(G)P_i(H)}{P_i(H)} = P_i(G).$$

So even though both P_1 and P_2 assign probability $1/2$ to A initially, learning that the coin lands heads yields $P_1(A|H) = 0.1$ and $P_2(A|H) = 0.9$. Hence, $P_1(A|H) < P_1(A) \leq P_2(A) < P_2(A|H)$. $\triangle$

Example 2.2 is a striking case. In a single step, learning the same evidence can transform agreement into significant disagreement. Moreover, the example points to an interesting and important connection between polarization, a topic in social epistemology, and dilation, a topic in the theory of individual imprecise probabilities. We turn to this connection now.

2.3 Dilation

The polarization phenomenon with which we were concerned in Section 2.2 bears resemblance to the phenomenon known as *dilation* in the theory of imprecise probabilities (IP). In fact, Example 2.2 is a borrowed example of dilation. In our opinion, there is a very fruitful exchange of ideas between social epistemology and the theory of imprecise probabilities (see, e.g., Levi, 1982, 1985a; Seidenfeld et al., 1989, 2010; Elkin and Wheeler, 2018; Stewart and Ojea Quintana, 2018).

IP allows for more general representations of uncertainty than standard Bayesian probability theory does. Several frameworks have been used to accomplish this task. We will consider *sets*

of probabilities, a very general IP framework. Let $\mathbb{P}$ denote a set of probability measures defined on the same measurable space $(\Omega, \mathcal{F})$. Levi marks an important distinction between two possible interpretations of $\mathbb{P}$ (1985b). On the one hand, we might retain the standard Bayesian ideal according to which any rational state of uncertainty admits representation in terms of a single probability function. On this account, a set $\mathbb{P}$ could be used to represent the possible values that an agent's precise probability judgments may take. Such a set might arise from failures of introspection or partial elicitation. In short, $\mathbb{P}$ may represent imprecision in measuring a precise credal state. For example, in estimating the probability of some event, an agent may be in a position to specify no more than one or two decimal places. On the other hand, IP can be seen as offering an alternative (laxer) normative standard. On this account, a set $\mathbb{P}$ might *completely describe* an agent's probability judgments in cases in which her uncertainty is not reducible to a unique probability function. Levi calls this interpretation *indeterminate* probability. Of course, both factors may operate simultaneously. There could be partial elicitation of an indeterminate state of uncertainty.

Dilation occurs, roughly speaking, when learning *increases* uncertainty. There are various formulations of dilation, depending on the choice of IP representation and whether certain inequalities are strict or not. But to facilitate comparison with the notion of polarization defined in the previous section, consider the following common definition.

Definition 2.2. Let $\mathbb{P}$ be a set of probabilities on $(\Omega, \mathcal{F})$, let $\mathscr{B}$ be a positive measurable partition of Ω[8], and let $A \in \mathcal{F}$. We say that the partition $\mathscr{B}$ *dilates* A just in case, for each $E \in \mathscr{B}$,

$$\inf\{P(A|E) : P \in \mathbb{P}\} < \inf\{P(A) : P \in \mathbb{P}\} \leq \sup\{P(A) : P \in \mathbb{P}\} < \sup\{P(A|E) : P \in \mathbb{P}\}.$$

The mathematics of dilation has been extensively studied in a series of articles (Seidenfeld and Wasserman, 1993; Herron et al., 1994; Wasserman and Seidenfeld, 1994; Herron et al., 1997; Pedersen and Wheeler, 2014, 2015). For results related to our Theorem 2.1, see in particular Wasserman and Seidenfeld's Result 1 (1994), and Pedersen and Wheeler's Theorem 1 and its corollary (2015).

Dilation is like a virulent form of polarization. In the social setting, it amounts to agents' views getting further from consensus no matter what outcome of a partition $\mathscr{B}$ they observe. In

[8]The partition $\mathscr{B}$ is positive and measurable if $E \in \mathscr{B}$ implies $E \in \mathcal{F}$ and $P(E) > 0$.

Example 2.2, where the partition is given by the outcomes of a toss of a fair coin ($\mathscr{B} = \{H, H^c\}$), P_1 and P_2 will move the same significant distance from agreement on A whether the coin lands heads or tails. It is the fact that a more precise estimate for an event A is transformed into a less precise estimate *regardless of the event in the partition that is learned* that attracts interest to dilation. While dilation emerges sometimes for IP, precise credal states are immune to dilation since $\inf\{P(A|E)\} = \sup\{P(A|E)\}$ for all A and E in $\mathcal{F}$.

Some see dilation as a pathological feature that calls IP into question (e.g., White, 2010). While not central to the present study, the debate about the normative status of dilation, we suggest, might be further illuminated by consideration of dilation in a social setting. Similarly, another line of inquiry into polarization would be to explore whether arguments against the rational acceptability of dilation can be extended in some way to social settings. Does the (un)acceptability of dilation provide some defeasible consideration in favor of the (un)acceptability of (certain kinds of) polarization or *vice versa*? Or is there, contrary to the hopes of those seeking decision theories and theories of inquiry unified across individuals and groups, some significant disanalogy?

2.4 Global Polarization

In the previous sections, our focus was a local sense of polarization. We were interested in situations in which sharing evidence increases the extent of disagreement between two Bayesian agents *with respect to a fixed event*. In some situations, however, there may not be a distinguished event around which inquiry centers, or one may be interested in whether a consensus can be reached about a whole collection of events. In such situations, a more global perspective is appropriate. In climate science, for example, there is a complex cluster of issues—from future sea levels and air temperature to the mechanisms underwriting decadal climate variability—that exercise researchers in the area. In some cases, consensus on appropriate measures to mitigate change or on adaptive policies might require consensus on a large number of other issues.

In order to adopt the global perspective, we need a way of measuring the extent to which two probabilities disagree that does not depend on a distinguished event. We will use the *total*

variational distance d defined for any probabilities P_1 and P_2 by

$$d(P_1, P_2) := \sup_{A \in \mathcal{F}} |P_1(A) - P_2(A)|.$$

Note that if P_1 and P_2 are in complete agreement, in the sense that $P_1(A) = P_2(A)$ for all events A, then the total variational distance between them is 0. If, on the other hand, P_1 and P_2 disagree *maximally*, in the sense that there's an event A such that $P_1(A) = 0$ and $P_2(A) = 1$, then $d(P_1, P_2) = 1$. The total variational distance finds use throughout probability theory and, as we explain in the next sections, has played a crucial role in Bayesian thought *via* the Blackwell-Dubins merging of opinions theorem (1962) and related results (Schervish and Seidenfeld, 1990; Huttegger, 2015b). In the examples below, we will make use of the fact that in finite probability spaces the total variational distance is given by

$$d(P_1, P_2) = P_1(A_0) - P_2(A_0), \tag{2.1}$$

where A_0 is the set of points $\omega \in \Omega$ such that $P_1(\omega) > P_2(\omega)$.[9] Using total variational distance, we can now introduce a notion of polarization that is the global analogue of polarization with respect to an event, as defined in Definition 2.1.

Definition 2.3. We say that evidence E *polarizes* P_1 *and* P_2 *globally* if $d(P_1, P_2) < d(P_1^E, P_2^E)$.

Note that the notion of polarization in Definition 2.3 does not depend on a particular event, but rather is concerned with the effect that learning has on the *overall* disagreement between two probabilities.

Does global polarization imply irrationality? Just as with polarization with respect to an event, the answer is no when the standard of rationality is Bayesian. This can be seen by considering a

[9] *Proof.* Note that A_0^c is the set of points $\omega \in \Omega$ such that $P_2(\omega) \geq P_1(\omega)$. For all $A \in \mathcal{F}$ we have

$$\begin{aligned} P_1(A) - P_2(A) &= \sum_{\omega \in A} \big[P_1(\omega) - P_2(\omega)\big] = \sum_{\omega \in A \cap A_0} [P_1(\omega) - P_2(\omega)] + \sum_{\omega \in A \cap A_0^c} [P_1(\omega) - P_2(\omega)] \\ &\leq \sum_{\omega \in A \cap A_0} [P_1(\omega) - P_2(\omega)] \leq \sum_{\omega \in A_0} [P_1(\omega) - P_2(\omega)] = P_1(A_0) - P_2(A_0). \end{aligned}$$

Similarly, $P_2(A) - P_1(A) \leq P_2(A_0^c) - P_1(A_0^c)$ for all $A \in \mathcal{F}$. Since $P_1(A_0) + P_1(A_0^c) = 1 = P_2(A_0) + P_2(A_0^c)$, we have $P_1(A_0) - P_2(A_0) = P_2(A_0^c) - P_1(A_0^c)$, and it follows from the above that $|P_1(A) - P_2(A)| \leq P_1(A_0) - P_2(A_0)$ for all $A \in \mathcal{F}$. The equality in (2.1) now follows by taking the supremum (maximum) of the left-hand side of the last inequality. □

slight variation of Example 2.1.

Example 2.3. *Let* $(\Omega, \mathcal{F})$*,* $A = \{\omega_1, \omega_2\}$*, and* $E = \{\omega_1, \omega_3\}$ *be defined as in Example 2.1, and consider the following priors and posteriors.*

Table 2.3

	ω_1	ω_2	ω_3	ω_4
P_1	1/4	1/8	1/2	1/8
P_2	1/2	1/12	1/4	1/6
P_1^E	1/3	0	2/3	0
P_2^E	2/3	0	1/3	0

Using (2.1), we can see that we have global polarization because $d(P_1, P_2) = 7/24 < 1/3 = d(P_1^E, P_2^E)$*. Note also that we still have polarization with respect to A, like in Example 2.1, because* $P_1(A \mid E) = 1/3 < 3/8 = P_1(A) \leq P_2(A) = 7/12 < 2/3 = P_2(A \mid E)$*.* △

Absent reason to believe that the above probability assignments are unreasonable, Example 2.3 shows that TOTAL is false under a global interpretation of "disagreements." In fact, as was the case in Section 2.2, the present example falsifies a thesis that is even weaker than TOTAL because global polarization, like polarization with respect to an event, implies persistent disagreement.[10]

Having defined two notions of polarization, it is natural to ask whether there are any interesting logical relations between them. Example 2.3 shows that both kinds of polarization can occur simultaneously. But does one notion imply the other? To begin to address this question, we return to Example 2.1. In that example, although E polarizes P_1 and P_2 with respect to A, it is not the case that E polarizes P_1 and P_2 globally. One can see this by using (2.1) to calculate $d(P_1, P_2) = 1/3 = d(P_1^E, P_2^E)$. By altering the probabilities in Example 2.1 a bit, one can find even more striking cases, in which there is polarization with respect to an event even though conditionalizing *decreases* the total variational distance between posteriors. For completeness, we

[10]If $d(P_1, P_2) < d(P_1^E, P_2^E)$, then $d(P_1^E, P_2^E) > 0$. This implies that there is some event A about which P_1^E and P_2^E disagree, i.e. $P_1^E(A) \neq P_2^E(A)$.

have included such an example in the Appendix (Example 2.4).

So we cannot infer global polarization from polarization with respect to an event. How about the converse implication? If evidence E polarizes two probabilities globally, does it follow that E polarizes the two probabilities with respect to some event? Again the answer is negative by another slight modification of the previous examples. See Example 2.5 in the Appendix. We summarize the previous conclusions with the following proposition.

Proposition 2.1. *There are cases in which E polarizes P_1 and P_2 both globally and with respect to some event A. However, the following implications do* ***not*** *hold.*

(i) If E polarizes P_1 and P_2 with respect to some event A, then E polarizes P_1 and P_2 globally.

(ii) If E polarizes P_1 and P_2 globally, then E polarizes P_1 and P_2 with respect to some event A.

Having established that global polarization and polarization with respect to an event are logically independent, it would be convenient to state a simple characterization of global polarization along the lines of Theorem 2.1. Although there are various methods for computing and comparing the total variational distances between posteriors and priors, we have not discovered a method that is sufficiently simple and illuminating for the purposes of this paper. Rather than introduce more technical material for relatively little philosophical payoff, we prefer to leave the task of finding a simple characterization of global polarization as an open problem. In future work, we plan to investigate global polarization in the context of imprecise probabilities. Global polarization gives rise to a phenomenon that is similar to dilation in some ways, but, like local and global polarization for precise probabilities, this phenomenon is logically independent of dilation for imprecise probabilities.

The important point for our purposes is that global polarization does not imply irrationality. We have now shown two senses in which TOTAL fails when the standard of rationality is Bayesian. Bayesian agents who learn a finite amount of shared evidence can exhibit both local polarization with respect to a distinguished event and global polarization with respect to their entire probability distributions. Whether we interpret TOTAL as requesting local or global resolutions of disagreement, we find that the thesis is false.

Yet, faith in the ability of rationality and evidence to avoid polarization may remain. Sure, learning just *one* event (or finitely many for that matter) allows for polarization. But doesn't

ongoing inquiry that allows for as many observations as we please avoid it? We show in Section 2.6 that, in fact, it does not. Taking inquiry to the limit does not save TOTAL.[11] But it is worth pointing out that retreating to the limit of inquiry already drastically weakens any automatic inference from the mere fact of actual polarization to the irrationality of some polarized agent or other. Presumably, all actual behavior occurs in the context of just finitely many observations. In such a context, this line of response concedes that polarization is consistent with the rationality of both parties.

Part II: The General Case

In the general setting, we consider cases of learning infinite amounts of evidence. Why bother looking at such artificial learning scenarios? In a way, we are pursuing TOTAL to its last retreat. Even allowing rational agents to learn an infinite amount of evidence, agreement cannot be ensured. Moreover, such learning scenarios are frequently considered in the context of Bayesian foundations. It is sometimes thought that we can test the mettle of a learning method by looking at its asymptotic behavior. Bengt Autzen gives recent voice to this idea. "Under the ideal scenario of an infinitely large data set," he writes, "an inference procedure should show certain desirable features" (Autzen, 2018, p. 3). One desirable feature claimed for Bayesian methodology is known as *convergence to the truth*. Convergence has long been taken to be a metric by which to judge probabilistic inference methods (e.g., Reichenbach, 1938, §43). Our concern in this paper, though, is with consensus. Consensus is no less pedigreed a methodological concern than convergence is. In "The Fixation of Belief," Peirce puts forward a picture of scientific method according to which a community of inquirers achieves consensus eventually by updating on shared evidence. As he puts it, "the method must be such that the ultimate conclusion of every man shall be the same" (Peirce, 1992a, p. 120).[12]

[11] In a sense, this is already clear even in the finite case. While Example 2.2 does not depend on extreme assignments, it could be adapted so that $P_1(G) = 0$ and $P_2(G) = 1$. In this case, evidence H maximally "polarizes" P_1 and P_2 with respect to A. And since 0 and 1 probabilities are not revisable under Bayesian conditionalization, such polarization is permanent. So there is no hope of undoing polarization or resolving disagreements concerning A for P_1 and P_2. Note that A is still *not* an event for which either prior is extreme in this modified example. So the point is not merely that achieving consensus is frustrated for those events with prior 0-1 assignments. In any case, disagreeing on prior 0-1 probabilities does not count as polarization as we define it, even if it is a case of interminable disagreement for Bayesians.

[12] Otherwise, a method will fail to fix belief because we will encounter those who disagree with us and, due to our "social impulse," our confidence will be shaken. In other places, Peirce seems to identify truth and whatever opinion the community settles on in the limit (e.g., 1992b).

Here, too, many Bayesians claim success. In the next section, we explain the basis for this claim. In short, then, we examine the general case because proponents of TOTAL may be tempted to appeal to such idealized scenarios and because such scenarios are standardly studied in Bayesian theory and even play foundational roles in justifications for the Bayesian point of view.

Before we can state the relevant facts about convergence and consensus for Bayesians, we need to bring a bit more machinery online. This machinery will help us in this second part of our case against TOTAL. As above, let $(\Omega, \mathcal{F}, P)$ be a probability space. We will be interested in situations in which agents anticipate learning an increasing amount of evidence that eventually settles every event of interest to them. The evidence is represented as a sequence of finite partitions, $\{\mathcal{E}_n\}_{n\in\mathbb{N}}$, such that $\mathcal{E}_{n+1}$ refines $\mathcal{E}_n$ for all $n \in \mathbb{N}$.[13] Because of the assumption that the partitions are increasingly fine, we say that the agent's evidence is *increasing*. For example, a partition might represent the possible outcomes of an experiment that the agent plans to perform. In the case of repeated coin tosses, $\mathcal{E}_1$ represents the information about the first toss of the coin, while $\mathcal{E}_2$ would represent all of the information about the first two tosses, etc. So by the time the agent observes the outcome of the second coin toss, she knows whether the "actual" sequence is one that begins HH or not.

We will also assume that the observations eventually settle every event in $\mathcal{F}$ and say that the evidence is *complete*. Formally, we require that the collection of all evidential events, namely $\bigcup_n \mathcal{E}_n$, *generate* the sigma-algebra $\mathcal{F}$ on which the agent's prior is defined. That is, we will assume that $\mathcal{F}$ is the smallest sigma-algebra containing $\bigcup_n \mathcal{E}_n$.

From her prior perspective, the agent is uncertain, for all n, which event in $\mathcal{E}_n$ she will learn. If $\omega \in \Omega$ is the actual world, then $E_n(\omega)$ denotes the event in $\mathcal{E}_n$ that the agent learns at stage n, namely, whichever member of $\mathcal{E}_n$ contains ω. In this setup, a Bayesian agent's posteriors are then $P(\cdot \mid E_n(\omega))$ for all $\omega \in \Omega$ and $n \in \mathbb{N}$.[14] The aforementioned *convergence to the truth* theorem says that a Bayesian agent assigns probability 1 to the event that her posterior probabilities converge to the truth about every event in $\mathcal{F}$. This means that for every $A \in \mathcal{F}$ and every $\omega \in \Omega$ in a set with P-probability 1, if A is true, so that $\omega \in A$, a Bayesian's posteriors $P(A \mid E_n(\omega))$ will

[13]The finite partition assumption aids exposition but is not necessary. In the Appendix, we relax it and work with general filtrations of sub-sigma-algebras.

[14]Technically, we may have $P(E_n(\omega)) = 0$ for some n and ω, in which case we can define $P(\cdot \mid E_n(\omega))$ arbitrarily and replace "for all $\omega \in \Omega$" with "for all $\omega \in \Omega$ in a set with P-probability 1."

get arbitrarily close to 1 as n increases. If A is false, so that $\omega \notin A$, then $P(A \mid E_n(\omega))$ will get arbitrarily close to 0 as n increases. See the Appendix for a more formal summary. We return now to our primary focus, consensus or the lack thereof.

2.5 Merging of Opinions

Merging of opinions is an important part of Bayesian lore. Relative to just a few assumptions, with probability 1, opinions get closer and closer together as they learn from a shared, increasing stream of data. Huttegger sees merging results as evidence that Bayesianism fulfills Peirce's vision of a method that settles belief for a community on the basis of experimental evidence. He writes, "experience trumps any initial belief state; diverging opinions are just a sign that not enough evidence has accumulated yet" (2015b, p. 613). Savage (1972), Blackwell and Dubins (1962), and Gaifman and Snir (1982) provide classic versions of merging of opinions theorems, which have since been generalized in various ways (Schervish and Seidenfeld, 1990; Huttegger, 2015b; Stewart and Nielsen, 2019). Such classic versions of these results attained their prominent theoretical status due to the subjective nature of personal probabilities. According to many Bayesians, that agents reach consensus (almost surely, relative to the assumptions in the theorems) should allay concerns that Bayesianism robs science of any sort of objectivity. "This approximate merging of initially divergent opinions is, we think, one reason why empirical research is called 'objective'," write Edwards, Lindman, and Savage (1963, p. 197).

Let P and Q be two probability measures on $(\Omega, \mathcal{F})$. Above, we explained that if P anticipates learning increasing and complete evidence, then P assigns probability 1 to converging to the truth. In this section, we continue to assume that evidence is increasing and complete, but we now also assume that it is *shared* with Q. We say that P *shares evidence with* Q if for all $n \in \mathbb{N}$ and $E \in \mathcal{E}_n$, if $P(E) > 0$, then $Q(E) > 0$. This ensures that Q can conditionalize on any evidential event that P can—anything that P can learn, Q can learn, too.

For all $\omega \in \Omega$, let $P_n(\omega) = P(\cdot \mid E_n(\omega))$ and $Q_n(\omega) = Q(\cdot \mid E_n(\omega))$ be the posteriors for P and Q after conditionalizing on a member of the n^{th} partition. As an informal gloss on the merging of opinions results, we might say, if P is "sufficiently similar" to Q, then for all ω in a set with P-probability 1 the distance between the posteriors $P_n(\omega)$ and $Q_n(\omega)$ goes to 0 as n increases. We

will use the same notion of distance that we used in Section 2.4, namely total variational distance. We say that P and Q *merge* if $d(P_n(\omega), Q_n(\omega))$ gets arbitrarily close to 0 for all ω as n increases, and we say that P *expects to merge with* Q if this event occurs for all ω in a set with P-probability 1.

We now need to say what we mean by sufficient similarity. Call P *absolutely continuous* with respect to Q when $Q(A) = 0$ implies $P(A) = 0$ for all $A \in \mathcal{F}$. In other words, any extreme probability assignment of Q's is an extreme probability assignment of P's. It could still be the case, though, that $P(A) = 0$ but $Q(A) > 0$ for some $A \in \mathcal{F}$. If Q is absolutely continuous with respect to P also, then we say that P and Q are *mutually absolutely continuous*.

Theorem 2.2 (Blackwell and Dubins (1962))**.** *Suppose that P shares increasing and complete evidence with Q. If P is absolutely continuous with respect to Q, then P expects to merge with Q.*

Given that P is absolutely continuous with respect to Q, P assigns probability 1 to approaching consensus with Q when they share increasing and complete evidence. If, in addition, Q is absolutely continuous with respect to P, then both P and Q expect to merge with each other.

But wait. Doesn't Theorem 2.2 show that polarization is essentially inconsistent with Bayesian rationality? Some authors do indeed seem to think polarization is beyond the pale, not just for some vague theory of rationality in general, but for Bayesianism in particular.

> Ample psychological evidence suggests that people's learning behavior is often prone to a "myside bias" or "irrational belief persistence" in contrast to learning behavior exclusively based on objective data. In the context of Bayesian learning such a bias may result in diverging posterior beliefs and attitude polarization even if agents receive identical information. Such patterns cannot be explained by the standard model of rational Bayesian learning that implies convergent beliefs. (Zimper and Ludwig, 2009, p. 181)

At least two points about Theorem 2.2 require careful consideration. First, the theorem has preconditions. To see that they are significant, note that Theorem 2.2 has a partial converse, which follows from the Bayesian Consensus-or-Polarization Law in the next part of the paper. It turns out that if P is *not* absolutely continuous with respect to Q, then either P does not expect to

merge with Q or P does not share evidence with Q. So, provided P shares evidence with Q, if P is *not* absolutely continuous with respect to Q, then P assigns positive probability to the event that the posteriors P_n and Q_n persist in disagreeing despite access to increasing and complete evidence. In section 2.6, we return to these issues with more precision. But we should pause now to think about the absolute continuity assumption.

Absolute continuity is not a *rationality* requirement. After all, with respect to which measure or measures ought a prior be absolutely continuous? Unless we have a principled answer to that question, it is difficult to even make sense of a proposal according to which absolute continuity is a normative constraint. And there is no trivial answer to the question in general. In large measurable spaces[15], there is no measure with respect to which all measures are absolutely continuous. Of course, even if P and Q were both absolutely continuous with respect to a third, distinguished measure, neither need be absolutely continuous with respect to the other. It might be tempting to urge priors to avoid extreme assignments. For instance, *regular* probability measures—which assign positive probability to all non-empty events—enjoy wide-spread support among philosophical probabilists (Shimony, 1955; Lewis, 1980; Skyrms, 1995). However, in large enough probability spaces, such measures are impossible and extreme prior probability assignments unavoidable.[16]

Similarly, absolute continuity is of dubious *descriptive* value. According to Miller and Sanchirico, the condition is even "difficult to interpret behaviorally" (1999, p. 171). But our concern is that absolute continuity just *assumes* a great deal of agreement out of the gate. When is that much initial agreement actually realized? Perhaps one could maintain that communities of scientists often endorse statistical models sufficiently similar so as to be absolutely continuous. As a descriptive claim, however, such a view needs empirical support. Earman considers something even stronger.[17] Mutual absolute continuity may be in some sense *constitutive* of a scientific community: "it could be held that decisions on zero priors help to define scientific communities and that an account of

[15]Such spaces arise when considering random variables with continuous distributions, for example. A random variable representing the unknown duration of a prizefight would be such a quantity.

[16]"Perhaps the reason that the absolute continuity assumption has gained such currency in the literature is that it is so plausible in a finite, or even countable setting. Even the stronger assumption that both players regard each state as at least possible [they have regularity in mind here] seems attractive, since all this rules out is dogmatism. But it would be a mistake to carry this intuition into the necessarily uncountable setting that is relevant here: obviously, in this case some events must receive zero measure" (Miller and Sanchirico, 1999, p. 179).

[17]Earman himself regards the fanfare concerning merging of opinions with a good deal of skepticism.

scientific inference must be relativized to a community" (1992, p. 142). In a debate about the descriptive adequacy of absolute continuity, such a response would verge on question-begging. If the very definition of a scientific community implies absolute continuity of its members' opinions, then no serious debate about the descriptive status of absolute continuity in research communities remains to be had.

Another suggestion is that an important *conceptual* distinction about disagreement can be explicated in terms of absolute continuity.[18] The distinction is between disagreement and *radical* disagreement, with radical disagreement being modeled by failures of absolute continuity. First, we note that radical disagreement on this explication is not symmetric, which may be a questionable feature. One prior may be in radical disagreement with another that is only in modest disagreement with it. Second, failures of absolute continuity can arise even when opinions are, in one sense, intuitively "close," just as absolutely continuous priors can be "far apart." If $Q(A) = 0$ but $P(A) = 0.000001$, for example, then P is not absolutely continuous with respect to Q; while $P(A) = 0.99$ and $Q(A) = 0.01$ does not preclude absolute continuity. Third, in order to avoid being a mere relabeling of when absolute continuity holds or not, the distinction should draw on some well-motivated, independent account of radical disagreement which may not be forthcoming in light of the sort of example just given.

The second issue about Theorem 2.2 requiring special attention appears in the consequent rather than the antecedent. P merges with Q *almost surely*. The distinction between *sure* and *almost sure* is not always kept firmly in mind when it comes to Bayesian convergence and merging results, leading some to make remarks apparently inconsistent with our general claim in this paper. More cynically, Earman writes, "'almost surely' sometimes serves as a rug under which some unpleasant facts are swept" (Earman, 1992, p. 148).[19] We may be interested in what holds surely (for all $\omega \in \Omega$), and not in what is merely highly probable.[20] Attending to the distinction between *merging*, on the one hand, and *merging almost surely*, on the other, we see that, if the "actual

[18]Thanks to an anonymous referee for this suggestion.

[19]Glymour (1980) raises a similar concern about convergence to the truth, writing, "The theorem does not tell us that in the limit any rational Bayesian will assign probability 1 to the true hypothesis and probability 0 to the rest; it only tells us that rational Bayesians are certain that he will" (p. 73). More recently, Gordon Belot has argued that convergence to the truth results constitute a liability for Bayesians because they forbid a "reasonable epistemological modesty" (2013, p. 502).

[20]See Nielsen (2018).

world" ω is outside the support of P, for all Theorem 2.2 says, P and Q may not *actually* merge. It is just that P assigns such points 0 probability. A number of examples have been used to motivate distinguishing certainty from 0-1 probability assignments. An agent might regard the tosses in an infinite sequence of fair coin tosses as independent. Then, any infinite sequence of (outcomes of) coin tosses has probability 0. Yet it would be a mistake to infer that she is certain that no such sequence is the actual one. An infinitely fine dart is thrown at the $[0, 1]$ interval. An agent's opinions about the outcome of the throw may be representable by the Lebesgue measure. Then, each real number in the unit interval bears probability 0. But the agent is not certain that, for each real number, the dart will not hit it.

Both the assumption of absolute continuity and the almost surely hedge require us to exercise considerable care in interpreting Blackwell and Dubins's theorem. In the remainder of the paper, we focus on absolute continuity. Even restricting our attention to probability "strong laws" (results that hold with probability 1), relaxing absolute continuity even to a very modest extent has significant ramifications.

2.6 The Bayesian Consensus-or-Polarization Law

The asymptotic analogue of polarization occurs when the total variational distance between posteriors tends to the maximal value of 1. More precisely,

Definition 2.4. P and Q *polarize in the limit* if $d(P_n(\omega), Q_n(\omega))$ gets arbitrarily close to 1 for all ω as n increases.

Polarization in the limit is basically stronger than the notion of global polarization in Definition 2.3. To explain this relation, suppose that P and Q polarize in the limit, and let r be the total variational distance between P and Q. It cannot be that $r = 0$ because then P and Q would be identical and would not be able to polarize in the limit. On the other hand, if $r = 1$, then polarization in the sense of Definition 2.3 can not occur. But this is a rather trivial limiting case. In cases of interest we may therefore assume that $0 < r < 1$. Since P and Q polarize in the limit, for all ω there is some stage n such that $d(P_n(\omega), Q_n(\omega)) > r$. It follows that $E_n(\omega)$ polarizes P and Q globally. Hence, excluding the trivial limiting case in which $r = 1$, if P and Q polarize in the limit, then for all ω there is some stage n such that $E_n(\omega)$ polarizes P and Q globally. It is in

this sense that polarization in the limit "basically" implies global polarization. Even if our notion of polarization in the limit is not the unique extension of Definition 2.3 to the general setting, it is a natural one in light of the relation just explained. Another reason to focus on it is that the total variational distance plays a prominent role in the theory of Bayesian merging, as evidenced by the common retort that "priors wash out."

In this section, we state and discuss a generalization of the Blackwell-Dubins merging of opinions result that we call the *Bayesian Consensus-or-Polarization Law*. To do this, we make use of a deep but easily explained result in measure theory called the Lebesgue Decomposition Theorem. In order to explain that result, we need one more definition. Let P and Q be any two probability measures. We say that P and Q are *mutually singular* if P assigns probability 1 to an event to which Q assigns probability 0.[21] When P and Q are mutually singular, absolute continuity fails in the most radical way possible. In subjective terms, P is probabilistically certain that some event A will occur while Q is certain that it will not. It follows that when P and Q are mutually singular, the total variational distance between them takes its maximal value of 1.

Now, for any two probability measures P and Q, the Lebesgue Decomposition Theorem implies that for some $\delta \in [0, 1]$, P can be decomposed as

$$P = \delta P^a + (1 - \delta)P^s, \tag{2.2}$$

where P^a is a probability measure that is absolutely continuous with respect to Q and P^s is a probability measure that is mutually singular with Q. Furthermore, if δ is strictly between 0 and 1, then the decomposition is unique. One can check that P is absolutely continuous with respect to Q if and only if $\delta = 1$. And similarly, P and Q are mutually singular if and only if $\delta = 0$.

The decomposition given by equation (2.2) tells us that P can be viewed as a mixture of two probabilities, one of which is absolutely continuous with respect to Q and one of which is mutually singular with Q. The larger δ is the "more" absolutely continuous P is with respect to Q. For this reason, we will call δ the *degree of absolute continuity* of P with respect to Q, and we say that P is absolutely continuous with respect to Q *to degree* δ. In view of our earlier remarks, P is absolutely

[21] Mutually singular probabilities have already made a brief appearance in our study of polarization. See footnote 11.

continuous with respect to Q *simpliciter* just in case P is absolutely continuous with respect to Q to degree 1.

The main result of this part of the paper demonstrates a tight connection between degree of absolute continuity and merging of opinions and generalizes the Blackwell-Dubins theorem that we discussed in the previous section.

Theorem 2.3 (Bayesian Consensus-or-Polarization Law)**.** *Let P and Q be any two probabilities and suppose that P is absolutely continuous with respect to Q to degree δ. Let M be the event that P and Q merge, and let L be the event that P and Q polarize in the limit. If P shares an increasing and complete sequence of evidence with Q, then*

$$P(M) = \delta \quad \textit{and} \quad P(L) = 1 - \delta.$$

Theorem 2.3 implies that an agent with probabilities given by P is certain, with probability 1, that either he and Q will merge or they will polarize: $P(M \cup L) = 1$. Under the assumption of shared, increasing, and complete evidence, it is incoherent for P to assign positive probability to the event that the distance between the posteriors P_n and Q_n tends to 0.6, for example, or that the distance oscillates forever. To see that our result generalizes the Blackwell-Dubins theorem, suppose that P is absolutely continuous with respect to Q. Then $\delta = 1$, so $P(M) = 1$, which is the conclusion of Theorem 2.2.

The idea of the proof of Theorem 2.3 is straightforward. Think of P as being determined by flipping a coin with bias δ. With probability δ, P is absolutely continuous with respect to Q and equal to P^a. With probability $1 - \delta$, P and Q are mutually singular and P is equal to P^s. If P ends up equal to P^a, then the Blackwell-Dubins theorem tells us that P expects to merge with Q. Hence, M occurs with probability δ, which is the first conclusion of our result. If, on the other hand, P ends up equal to P^s, then $P(A) = 1$ and $Q(A) = 0$ for some event A. No amount of conditionalizing can change these extreme probability assignments, so P and Q must polarize in the limit. Hence, L occurs with probability $1 - \delta$. Making this argument precise requires some careful management of probability 0 events, which is where the assumption that evidence is shared comes into play. The complete proof can be found in the Appendix.

Theorem 2.3 also implies the converse to the Blackwell-Dubins theorem that was advertised in

the last section. Suppose that P shares increasing and complete evidence with Q and that P expects to merge with Q. Then $1 = P(M) = \delta$ by Theorem 2.3. So P must be absolutely continuous with respect to Q. We see that in the presence of shared, increasing, and complete evidence absolute continuity is necessary for merging of opinions to occur with probability 1. We record this fact as the following corollary.

Corollary 2.1. *If P shares an increasing and complete sequence of evidence with Q, and P expects to merge with Q, then P is absolutely continuous with respect to Q.*[22]

One thing that we would like to stress about Theorem 2.3 is that the way in which we relax absolute continuity is quite mild because we continue to assume that the absolute continuity relation holds with respect to evidential events. Our assumption is that for any event E that can be expressed as $E = E_1 \star ... \star E_k$ (where each E_i is an element of some partition $\mathcal{E}_n$ and $\star$ is some operation on sets), that is, any event that can be settled by a finite amount of evidence, $P(E) = 0$ if $Q(E) = 0$. In other words, absolute continuity is relaxed just for infinitary events. Relaxing absolute continuity *further* than we have would be to relinquish the shared evidence assumption: there may be events that P can learn at some stage n that Q cannot. So, our assumption is a motivated way to mildly relax absolute continuity. However, the conclusion of our theorem is importantly different from that of Blackwell and Dubins's. This suggests to us that the classic merging of opinions results, by failing to be robust even to our mild weakening of the assumptions, are something of an artifact of the under-motivated but strong assumption of absolute continuity.

2.7 Discussion

About polarization, Thomas Kelly asks, "Given that You and I are responding to our evidence in such-and-such a way, is there any chance that our doing so is anything other than blatantly unreasonable?" (2008, p. 631). More generally, are there grounds to deny TOTAL? Our study here provides an affirmative answer when the standard of reasonableness is Bayesian. As we have shown, it is trivial to find instances of polarization and persistent disagreement when agents learn just a finite amount of shared evidence. In the general case, the guarantee of asymptotic consensus for

[22]We note that this result was first proved by Kalai and Lehrer (1994). In the Appendix we actually generalize their result because their proof assumes that evidence is represented by partitions while ours does not.

Bayesians is an artifact of special auxiliary assumptions, like absolute continuity, that supplement Bayesian updating. Claims that rational learning leads to consensus require endowing the auxiliary assumptions with a normative status. We find no plausible channel for such an endowment.

Our investigation could be taken in two different ways. On the one hand, we could hold to the normative standard provided by Bayesian probability theory. On the other hand, we could deny that the standard Bayesian picture of learning presents us with a sound scientific methodology. In the former case, it seems that we must relinquish any requirement that sound scientific methodology provides investigators with resources to resolve disagreements through shared evidence. Again, somewhat confusingly, claims contrary to this view are routinely made from inside the Bayesian camp. Suppes, for example, writes,

> It is of fundamental importance to any deep appreciation of the Bayesian viewpoint to realize the particular form of the prior distribution expressing beliefs held before the experiment is conducted is not a crucial matter [...] For the Bayesian, concerned as he is to deal with the real world of ordinary and scientific experience, the existence of a systematic method for reaching agreement is important [...] The well-designed experiment is one that will swamp divergent prior distributions with the clarity and sharpness of its results, and thereby render insignificant the diversity of prior opinion (1966, p. 204).

Similar claims in the context of merging of opinions can be found even in the most recent literature.[23]

Consider now the second way to construe our study. There is more to scientific methodology, this reply goes, than is dreamt of in the Bayesian philosophy of coherence and conditionalization.[24] If we are to retain a probabilistic epistemology, we must constrain the class of "rationally permissible" priors rather substantially. But in a broadly Bayesian paradigm, the consensus requirement is tied to absolute continuity. Not only is absolute continuity a necessary condition for merging (whenever evidence is shared) that lacks normative motivation; it is a condition which we have no reason to expect to be satisfied in many cases. Furthermore, Theorem 2.2 only secures merging almost surely.

[23]See, for example, "We follow, e.g., Peirce in requiring that sound Scientific methodology provides investigators with the resources to resolve interpersonal disagreements through shared evidence" (Cisewski et al., 2018). But for which priors, with respect to how much evidence? And surely or only almost surely?

[24]We are not considering approaches that fully abandon Bayesianism or fully denounce its subjective elements.

Holding out hope that a case can be made for the normative status of the absolute continuity condition, one might think that perhaps there are other aspects of sound methodology that would imply absolute continuity, and thereby secure merging. What is the nature of these additional aspects of scientific methodology? Are they all agent-invariant, or are some of them more subjective? By focusing solely on revising judgments of subjective probability *via* conditionalization, we have made the case against TOTAL harder to make than it might have been—at least in one sense. Adding parameters that are *not* invariant across rational agents would only make achieving consensus a less likely outcome. According to Isaac Levi's epistemological outlook, to take an example with many subjective parameters, not only is what an agent "learns" or comes to accept determined in part by a subjective *value for information*, an agent can also revise her prior independently of learning new evidence in some circumstances (Levi, 1980).[25] So eliminating disagreements *via* rational learning is even less plausible on an account like Levi's. But what about additional aspects of good methodology that are objective? A popular objective sort of constraint on probability judgments is the Principal Principle. But it is difficult to see how such a principle would help. In general, the Principal Principle does not pin down a unique prior probability (if for no other reasons than that would require, first, too much shared "knowledge of the chances" of events and, second, that chances are reasonably attributable to *all* relevant events).[26] Typically, only small fragments of a distribution are determined by it. This, of course, leaves ample occasion for the failure of absolute continuity between different priors. So much the worse for TOTAL.

[25] In Levi's terminology, an agent can revise her *confirmational commitment*, which is a function from states of full belief to sets of probabilities, should circumstances call for it—without changing her state of full belief (see, e.g., Levi, 2009, §2).

[26] An anonymous referee points out that a result due to Deutsch in the context of the many-worlds interpretation of quantum mechanics can be interpreted as delivering a unique rational probability. See (Greaves, 2007, p. 115).

Appendix

Examples from Section 2.4. The following two examples were referred to in Section 2.4. The first is a case of polarization with respect to an event accompanied by a decrease in total variational distance.

Example 2.4. *Let $(\Omega, \mathcal{F})$, A, and E be defined as in Example 2.1, and consider the following priors and posteriors*

Table 2.4

	ω_1	ω_2	ω_3	ω_4
P_1	1/24	5/12	1/12	11/24
P_2	1/2	1/12	1/4	1/6
P_1^E	1/3	0	2/3	0
P_2^E	2/3	0	1/3	0

We still have polarization with respect to A because $P_1(A \mid E) = 1/3 < 11/24 = P_1(A) \leq P_2(A) = 13/24 < 2/3 = P_2(A \mid E)$. And yet, $d(P, Q) = 5/8 > 1/3 = d(P_1^E, P_2^E)$. $\triangle$

The next example is a case of global polarization without polarization with respect to any event.

Example 2.5. *Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$, $\mathcal{F} = 2^\Omega$, and $E = \{\omega_1, \omega_2\}$. Consider the priors and posteriors in the table below.*

Table 2.5

	ω_1	ω_2	ω_3
P_1	1/3	1/3	1/3
P_2	4/15	2/5	1/3
P_1^E	1/2	1/2	0
P_2^E	2/5	3/5	0

We have $d(P_1, P_2) = 1/15 < 1/10 = d(P_1^E, P_2^E)$, so E polarizes P_1 and P_2 globally. However, there is no subset of Ω with respect to which E polarizes P_1 and P_2. This is straightforward to verify and we omit the details. $\triangle$

Proof of Theorem 2.3. In the remainder of this appendix, we will provide a more formal and general presentation of the mathematical framework of Part 2. Then we will prove Theorem 2.3.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. We say that an event $A \in \mathcal{F}$ occurs *almost surely* with respect to P when $P(A) = 1$. We also say A occurs *a.s. (P)*, and if Q is another probability measure on $(\Omega, \mathcal{F})$ with respect to which A occurs almost surely, we say that A occurs a.s. (P/Q). We denote the indicator function for A by $\mathbf{1}_A : \Omega \to \{0, 1\}$, defined as

$$\mathbf{1}_A(\omega) = \begin{cases} 1, \text{ if } \omega \in A; \\ 0, \text{ otherwise.} \end{cases}$$

When $\mathcal{F}$ and $\mathcal{G}$ are both sigma-algebras of subsets of Ω, we call $\mathcal{G}$ a *sub-sigma-algebra* of $\mathcal{F}$ if $\mathcal{G} \subseteq \mathcal{F}$. Intuitively, sigma-algebras represent bodies of information. For example, consider tossing a coin N times. Prior to any tossing, one is uncertain what the actual sequence of tosses will be and has a prior distribution over all the binary sequences of length N. After observing the first toss one now knows whether the actual sequence begins with H or T. This information is represented by the sigma-algebra $\mathcal{G}$ that partitions all the possible binary sequences into those beginning with H and those beginning with T. Formally,

$$\mathcal{G} = \{\emptyset, \{\omega \in \Omega : \omega \text{ begins with } H\}, \{\omega \in \Omega : \omega \text{ begins with } T\}, \Omega\}.$$

If $\mathcal{F}$ is a sigma-algebra, then a real-valued function $X : \Omega \to \mathbb{R}$ is called *$\mathcal{F}$-measurable* if $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$ for every $x \in \mathbb{R}$. Intuitively, the function X is $\mathcal{F}$-measurable if every question about the values that X takes can be answered by the information in $\mathcal{F}$. To take a simple example, if $\mathcal{G}$ is defined as above and $A = \{\omega \in \Omega : \omega \text{ begins with } H\}$, then $\mathbf{1}_A$ is $\mathcal{G}$-measurable.

In the main text, we assumed for simplicity that evidence is represented by a sequence of finite partitions and that at each stage an agent learns an event in a partition. We now relax that assumption. We now assume that evidence is represented by a *filtration* $\{\mathcal{F}_n\}_{n \in \mathbb{N}}$ on $(\Omega, \mathcal{F})$, which

is defined to be a collection of sub-sigma-algebras of $\mathcal{F}$ such that $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$ for all $n \in \mathbb{N}$. To see that this generalizes the partition model used in the main text, note that every finite partition $\mathcal{E}_n$ generates a sub-sigma-algebra, the members of which are unions of members of $\mathcal{E}_n$ (as well as $\emptyset$). The inclusion relationship $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$ is the generalization of the above requirement that later partitions refine earlier ones and captures the idea that evidence is *increasing*. Let $\mathcal{F}_\infty$ be the smallest sigma-algebra containing $\bigcup_{n\in\mathbb{N}} \mathcal{F}_n$. In the general setting, we say that the filtration $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ is *complete* if $\mathcal{F}$ is generated by the sub-sigma algebras $\mathcal{F}_n$ in the sense that $\mathcal{F}_\infty = \mathcal{F}$. As before, the point of this assumption is that the evidential information contained in the filtration eventually captures all events of interest (events in $\mathcal{F}$).

Since evidence is represented by a filtration, we need a notion of conditional probability given a sub-sigma-algebra, whereas in the main text we were able to make do with the familiar notion of conditional probability given an event. We use the definition of conditional probability that is standard in modern probability theory. The conditional probability $P(A|\mathcal{F}_n)$ of A given the sub-sigma-algebra $\mathcal{F}_n$ is an $\mathcal{F}_n$-measurable function that satisfies $P(A \cap E) = \int_E P(A|\mathcal{F}_n)dP$ for all $E \in \mathcal{F}_n$. The existence of such a function is guaranteed for any sub-sigma-algebra by the Radon-Nikodym Theorem and is unique up to sets of P-measure 0. In other words, any $\mathcal{F}_n$-measurable function X that satisfies the given integral equation is almost surely equal to $P(A \mid \mathcal{F}_n)$ with respect to P, and we say that X is a *version* of $P(A \mid \mathcal{F}_n)$. Note that in the simple case where each $\mathcal{F}_n$ is generated by a finite partition $\mathcal{E}_n$ the conditional probabilities given $\mathcal{F}_n$ are given by

$$P(A \mid \mathcal{F}_n)(\omega) = P(A \mid E_n(\omega))$$

for all ω in a set with probability 1, which corresponds with the treatment of conditional probabilities in the main text.

The convergence-to-the-truth theorem, as stated in the main text, generalizes to

$$\lim_{n\to\infty} P(A|\mathcal{F}_n) = \mathbf{1}_A \text{ a.s.}(P)$$

for all $A \in \mathcal{F}$. As in the main text, all limits are pointwise in $\omega \in \Omega$. Another way of stating

convergence to the truth, then, is

$$P(\{\omega \in \Omega : \lim_{n\to\infty} P(A \mid \mathcal{F}_n)(\omega) = \mathbf{1}_A(\omega)\}) = 1$$

Following our previous notation, we write $P_n(\omega) = P(\cdot|\mathcal{F}_n)(\omega)$ and $Q_n(\omega) = Q(\cdot|\mathcal{F}_n)(\omega)$. In order to state the Blackwell-Dubins merging of opinions theorem in full generality, we need to make a further assumption about P_n and Q_n. For all that we have said so far, for fixed $\omega \in \Omega$, either or both of $P_n(\omega)$ and $Q_n(\omega)$ may fail to be probability measures on $(\Omega, \mathcal{F})$. If there are versions of P_n and Q_n such that $P_n(\omega)$ and $Q_n(\omega)$ are probabilities for all $\omega \in \Omega$, then we say that these versions are *regular* conditional probabilities. Under the assumption that $\mathcal{F}_n$ is generated by a finite partition, regular versions of P_n and Q_n exist. We can also guarantee the existence of regular versions of P_n and Q_n by making assumptions about the space $(\Omega, \mathcal{F})$. For example, if Ω is a Polish space and $\mathcal{F}$ is its Borel sigma-algebra, then regular versions of the conditional probabilities exist. But without further assumptions about the probability space or filtration, regular versions may not exist, so we now add their existence as a further assumption of the framework.[27]

If P is absolutely continuous with respect to Q, then we write $P \ll Q$. We say that P *shares evidence with* Q if $P|_{\mathcal{F}_n} \ll Q|_{\mathcal{F}_n}$ for all n, where $P|_{\mathcal{F}_n}$ and $Q|_{\mathcal{F}_n}$ denote the restrictions of P and Q to the sub-sigma-algebra $\mathcal{F}_n$. Note that this corresponds with the definition of sharing evidence in the main text when $\mathcal{F}_n$ is generated by a finite partition. We say that P and Q *merge* when

$$\lim_{n\to\infty} d(P_n, Q_n) = 0.$$

The general version of Theorem 2.2 is

Theorem 2.4 (Blackwell and Dubins (1962))**.** *Let P and Q be probability measures on $(\Omega, \mathcal{F})$, and let $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ be a complete filtration. Suppose that P_n and Q_n are regular conditional probabilities for all n. If $P \ll Q$, then P and Q merge with P-probability 1, i.e.*

$$P(\{\omega \in \Omega : \lim_{n\to\infty} d(P_n(\omega), Q_n(\omega)) = 0\}) = 1.$$

[27]For more on regular conditional probabilities, see Seidenfeld (2001) and Durrett (2010, 4.1c).

The general version of Theorem 2.3, which we are now ready to prove, is

Theorem 2.5 (Bayesian Consensus-or-Polarization Law, General Version)**.** *Let P and Q be two probability measures on $(\Omega, \mathcal{F})$, and let $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ be a complete filtration. Suppose that P_n and Q_n are regular conditional probabilities for all n, that P shares evidence with Q, and that P is absolutely continuous with respect to Q to degree δ. Then*

$$P(\{\omega \in \Omega : \lim_{n\to\infty} d(P_n(\omega), Q_n(\omega)) = 0\}) = \delta$$

and

$$P(\{\omega \in \Omega : \lim_{n\to\infty} d(P_n(\omega), Q_n(\omega)) = 1\}) = 1 - \delta.$$

Proof. If $\delta = 1$, then the result follows from Theorem 2.4, so assume $\delta < 1$. Let the Lebesgue decomposition of P with respect to Q be given by

$$P = \delta P^a + (1 - \delta)P^s,$$

as in equation (2.2) of the main text. Clearly, $P^a \ll P$ and $P^s \ll P$.

By the triangle inequality, $P^a \ll Q$, $P^a \ll P$, and Theorem 2.4,

$$d(P_n, Q_n) \leq d(P_n, P_n^a) + d(P_n^a, Q_n) \to 0 \text{ a.s. } (P^a) \tag{2.3}$$

With $M = \{\omega \in \Omega : d(P_n(\omega), Q_n(\omega)) \to 0\}$, (2.3) implies

$$P^a(M) = 1. \tag{2.4}$$

Next, let $A^s \in \mathcal{F}$ be such that $Q(A^s) = 0$ and $P^s(A^s) = 1$. Such an event exists because P^s and Q are mutually singular. Since $Q(A^s) = 0$, we have $Q_n(A^s) = 0$ a.s. $(Q/Q|_{\mathcal{F}_n})$ since Q_n is $\mathcal{F}_n$-measurable. This implies $Q_n(A^s) = 0$ a.s. $(P|_{\mathcal{F}_n}/P)$, which then implies $Q_n(A^s) = 0$ a.s. (P^s). Since $P^s(A^s) = 1$, we have $P_n^s(A^s) = 1$ a.s. (P^s). In this way, for all n we find a P^s-probability 1 set on which $Q_n(A^s) = 0$ and $P_n^s(A^s) = 1$. It follows that $d(P_n^s, Q_n) = 1$ for all n a.s. (P^s). Using

this fact, the triangle inequality, $P^s \ll P$, and Theorem 2.4, we get

$$d(P_n, Q_n) \geq d(P_n^s, Q_n) - d(P_n^s, P_n) \to 1 \text{ a.s. } (P^s) \tag{2.5}$$

With $L = \{\omega \in \Omega : d(P_n(\omega), Q_n(\omega)) \to 1\}$, (2.5) implies

$$P^s(L) = 1. \tag{2.6}$$

Now, $M \subseteq L^c$ and $L \subseteq M^c$, so (2.4) and (2.6) imply

$$P^a(L) = 0 = P^s(M). \tag{2.7}$$

Using (2.4), (2.6), (2.7) and the Lebesgue decomposition (2.2) of P with respect to Q we compute

$$P(M) = \delta \quad \text{and} \quad P(L) = 1 - \delta,$$

which is the desired result. □

Chapter 3

Convergence to the Truth Without Countable Additivity

3.1 Introduction

Convergence results play an important role in the foundations of Bayesian epistemology and philosophy of science. For example, a standard reply to the worry that Bayesian theory is too subjective is that posterior probabilities converge to the truth with increasing evidence.[1] Underwriting this reply is a collection of mathematical theorems about the asymptotic behavior of conditional probabilities. The theorems typically cited in the philosophical literature are results in the measure-theoretic probability theory of Kolmogorov (1950) and make essential use of the countable additivity axiom. The first detailed application of convergence theorems to philosophical issues is due to Gaifman and Snir (1982). Their Theorem 2.1 makes the important distinction between (I) the posterior probabilities of an individual Bayesian agent converging to the truth, and (II) the posterior probabilities of multiple Bayesian agent's converging to a consensus. Although results concerning (II) were also shown by Blackwell and Dubins (1962), the Gaifman-Snir paper, whose results were proved independently of the Blackwell-Dubins result, was the first to draw a sharp philosophical distinction between (I) and (II) in a language-oriented framework that makes this distinction very natural. In both cases, however, countable additivity is crucial. But, although countable additivity is a powerful tool for proving mathematical theorems, it is a matter of considerable philosophical controversy whether it should be taken as an axiom regulating subjective probabilities. Some prominent theorists, the likes of which include de Finetti and Savage, have explicitly rejected the countable additivity axiom in favor of more general finitely additive theories of probability.[2] Given

[1] See Edwards et al. (1963).

[2] de Finetti (1972, 1974); Savage (1972).

the lack of consensus about countable additivity's status, it is natural to ask what can be said about the Bayesian convergence theorems when countable additivity is relaxed. The purpose of this paper is to initiate a systematic study of this question.

I will begin, in the next section, by giving an overview of the relevant mathematical and philosophical literature. Finitely additive convergence theorems have certainly been studied before, so one of my aims in the next section will be to compare and contrast the approach of the present paper with what has already been done. I will also relate the project to some recent developments in philosophy of science and formal epistemology. Having motivated the paper, I then introduce the mathematical framework in Section 3.3. The main results are in Section 3.4. Section 3.5 concludes by discussing some questions for future research.

3.2 Overview and Motivation

The results to follow are basically finitely additive martingale convergence theorems for conditional probabilities. In the countably additive setting, such results are usually traced back to Lévy (1937) and Doob (1953), though Ville (1936, 1939) is another early source of martingale theory.[3] Nowadays, martingale convergence theorems are an essential and standard part of measure-theoretic probability, which assumes countable additivity.[4]

There is much less work on the subject under the weaker assumption of finite additivity. This is partly due to the fact that there is not a single, widely-accepted theory of conditional probability in the finitely additive framework that is general enough to serve as a counterpart to the theory of conditional probability that Kolmogorov developed under countable additivity. Still, some efforts have been made. In mathematics, the finitely additive framework that has proved most fruitful for exploring convergence problems was introduced by Dubins and Savage (1965) using the notion of a *strategy*. Purves and Sudderth (1976), Chen (1977), and Purves and Sudderth (1983) have used the strategic framework to study almost sure convergence in some detail.[5] In view of this work, it is worth mentioning four differences between the results to follow and the results already proved in the strategic setting. First, the approach to finitely additive convergence that we develop

[3]See Shafer and Vovk (2005) and Bienvenu et al. (2009).

[4]For textbook treatments, see, e.g., Billingsley (2008) and Durrett (2010).

[5]For references to this work in the philosophical literature, see Zabell (2002); Skyrms (2006); Elga (2016).

here constitutes a minimal departure from the familiar measure-theoretic approach to probability theory. In order to study convergence in the strategic framework of Dubins and Savage, on the other hand, one must first develop quite a bit of new technical machinery. The results below should be accessible to anyone acquainted with standard, countably additive probability theory. For instance, the only notion of conditional probability that we use is the elementary one that defines conditional probabilities as ratios of unconditional probabilities. It should be noted that this is not without some loss of generality, however. This leads us to the second distinguishing feature of our framework: we focus on measurable spaces that are products of finite spaces. The strategic framework is more general in this regard as it applies to arbitrary product spaces. We have already noted that this requires non-trivial technical work. From a philosophical perspective, it is not clear that the generality afforded by this work is worth the effort. Our framework adequately models repeated experiments with finitely many outcomes. Unless one insists on experiments with infinitely many outcomes, our framework should be sufficient. Third, our main focus is not almost sure convergence but rather *almost uniform* convergence. In the countably additive setting, these two modes of convergence are equivalent, but in the finitely additive setting the latter is strictly stronger than the former. Almost uniform convergence is, arguably, a more natural mode of convergence to study in the finitely additive setting than almost sure convergence because, even under mere finite additivity, it implies convergence in probability; the same is not true of almost sure convergence. As far as I know, there is no precedent for studying almost uniform convergence in a finitely additive setting, so this aspect of the paper is something of a conceptual novelty. Fourth, and finally, whereas the work mentioned above focuses on conditions that are *sufficient* for finitely additive convergence, the main results here are *characterizations* that provide both necessary and sufficient conditions.

So much for the mathematical novelties. Why are the results below philosophically interesting? I will briefly mention two philosophical debates to which the results in this paper apply. One consideration sometimes given in favor of countable additivity is the mathematical fruit that it bears. The convergence theorems are a central example of countable additivity's fecundity. In view of this consideration, it is important to understand exactly when finite additivity is capable of delivering the same results. The theorems below shed some light on the situation.

The second relevant philosophical debate is a bit more subtle than the first. In recent work,

Gordon Belot (2013; 2017) has argued that convergence theorems constitute an epistemological *liability* for Bayesians. The theorems tell us that Bayesian agents are certain (with probability 1) that with increasing evidence their posteriors will converge to the truth for any hypothesis under consideration. But, Belot points out, there are hypotheses for which failure to converge to the truth is the *typical* outcome. For such hypotheses tempering one's certainty in convergence to the truth seems reasonable, but Bayesianism forbids this. Belot concludes that Bayesian theory is inconsistent with a "reasonable epistemological modesty" (Belot, 2013, p. 502). Al-Najjar et al. (2014) and Pomatto and Sandroni (2018) make a similar point in terms of *inductive skepticism*: in view of the convergence theorems, Bayesian agents are forbidden from entertaining skeptical worries of the sort developed by Hume (1748) and Goodman (1955).

Belot's argument has already prompted a number of replies.[6] The reply that is relevant for our purposes notes that Belot's argument is valid only under the assumption of countable additivity. Bayesian agents with merely finitely additive priors may assign positive probability to failing to converge to the truth (Elga, 2016). So, Bayesian theory without countable additivity is consistent with the sort of epistemological modesty that Belot finds reasonable. Similarly, merely finitely additive Bayesians can entertain Humean and Goodmanian skeptical doubts. What this line of response leaves open is the exact conditions under which Bayesian theory is consistent with modesty and inductive skepticism. The results to follow address this problem.

3.3 Preliminaries

This section has three subsections. The first contains all of the basic definitions and facts used throughout the rest of the paper. It is organized hierarchically: we begin by defining some set-theoretic structures, then we add some topological concepts, and finally we introduce probability measures and conditional probabilities. In the second subsection, we introduce three modes of convergence and explain the logical relations between them. The final subsection defines convergence to the truth.

[6] Huttegger (2015a); Weatherson (2015); Elga (2016); Cisewski et al. (2018); Pomatto and Sandroni (2018).

3.3.1 Basic Definitions

We will assume that the reader has some familiarity with elementary set theory, topology, and measure-theoretic probability. We omit definitions when our usage is standard.

Set Theory

If X is a set, then $\mathcal{A}$ is an *algebra* of subsets of X if $X \in \mathcal{A}$ and $\mathcal{A}$ is closed under finite unions and complementation. An algebra of subsets of X that is closed under countable unions is called a *sigma-algebra*. If $\mathcal{B} \subset 2^X$, then $\sigma(\mathcal{B})$ is the sigma-algebra *generated* by $\mathcal{B}$, the smallest sigma-algebra containing $\mathcal{B}$. Members of algebras are called *events*.

If X is a set and $\mathcal{A}$ is an algebra of subsets of X, then we call $(X, \mathcal{A})$ a *pre-measurable space*. If $\mathcal{A}$ is a sigma-algebra, then $(X, \mathcal{A})$ is called a *measurable space*.

If $\mathcal{A}$ is an algebra of subsets of X and $\mathcal{B} \subseteq \mathcal{A}$ is an algebra (resp. sigma-algebra) of subsets of X, then $\mathcal{B}$ is called a *sub-algebra (resp. sub-sigma-algebra)* of $\mathcal{A}$.

Let $(X, \mathcal{A})$ be a measurable space. A *filtration* of $(X, \mathcal{A})$ is a sequence $(\mathcal{A}_n)_{n\in\mathbb{N}}$ of sub-algebras of $\mathcal{A}$ such that $\mathcal{A}_1 \subseteq \mathcal{A}_2 \subseteq ...$ and $\mathcal{A} = \sigma(\bigcup_{n=1}^{\infty} \mathcal{A}_n)$. Note that if $(\mathcal{A}_n)_{n\in\mathbb{N}}$ is a filtration of $(X, \mathcal{A})$, then $\bigcup_{n=1}^{\infty} \mathcal{A}_n$ is an algebra of subsets of X.

Throughout the paper, we let $\Omega = \{0,1\}^{\mathbb{N}}$ be the set of all binary sequences. If $\omega \in \Omega$ and $n \in \mathbb{N}$, then $\omega(n) \in \{0,1\}$ denotes the nth coordinate of ω, and ω^n denotes the set of all binary sequences that agree with ω in the first n coordinates, i.e.

$$\omega^n =_{df} \{\omega(1)\} \times ... \times \{\omega(n)\} \times \{0,1\}^{\mathbb{N}}. \tag{3.1}$$

We let $\mathcal{C}$ denote the algebra of subsets of Ω generated by sets of the form in (3.1). If $C \in \mathcal{C}$, then there exist $n, m_1, ..., m_n \in \mathbb{N}$ and $\omega_1, ..., \omega_n \in \Omega$ such that $C = \omega_1^{m_1} \cup ... \cup \omega_n^{m_n}$. Let $\mathcal{F} = \sigma(\mathcal{C})$.

There is a natural filtration that generates $\mathcal{F}$. For each $n \in \mathbb{N}$, let

$$\mathcal{C}_n = \sigma\big(\{\omega^n : \omega \in \Omega\}\big). \tag{3.2}$$

Each member of $\mathcal{C}_n$ is a finite (possibly empty), pairwise disjoint union of sets of the form (3.1). Note that $|\mathcal{C}_n| = 2^n$ and $\mathcal{C} = \bigcup_{n=1}^{\infty} \mathcal{C}_n$.

If $(C_n)_{n\in\mathbb{N}}$ is a sequence of subsets of X, then the sequence of subsets of X defined by $D_n = C_n - \bigcup_{i=1}^{n-1} C_i$ for all $n \in \mathbb{N}$ is called the *disjoint counterpart* of $(C_n)_{n\in\mathbb{N}}$. One can easily verify that the members of $(D_n)_{n\in\mathbb{N}}$ are pairwise disjoint, that $\bigcup_{i=1}^{n} C_i = \bigcup_{i=1}^{n} D_i$ for all $n \in \mathbb{N} \cup \{\infty\}$.

Topology

We equip $\{0,1\}$ with the discrete topology and Ω with the product topology. The subsets of Ω of the form ω^n form a countable basis for the product topology and we call such sets *basic clopen sets*. Events in $\mathcal{C}$ are called *clopen*. In addition to being second countable, Ω is compact and completely metrizable.

The sigma-algebra $\mathcal{F}$ introduced in the previous subsection is the Borel sigma-algebra over Ω, the sigma-algebra generated by the basic clopen sets.

The members of the disjoint counterpart of a sequence of clopen subsets of Ω are themselves clopen. Thus, every open subset of Ω is the union of a countable, pairwise disjoint sequence of clopen sets.

Clopen events are epistemically significant because they are the events that can be confirmed or disconfirmed after sampling from Ω finitely many times. If $C \in \mathcal{C}$ is of the form $C = \omega_1^{m_1} \cup ... \cup \omega_n^{m_n}$, then, for any $\omega \in \Omega$, one can determine whether or not $\omega \in C$ by observing the first $\max_i\{m_i : 1 \leq i \leq n\}$ coordinates of ω.

Probability

Let $(X, \mathcal{A})$ be a pre-measurable space. A *probability measure* $P : \mathcal{A} \to [0,1]$ on $(\Omega, \mathcal{A})$ is a non-negative, finitely additive set function such that $P(X) = 1$. If P is also countably additive, then P is called a *countably additive probability measure*. If P is a probability measure that is not countably additive, then we say that P is *merely finitely additive*. If P is a (countably additive, merely finitely additive) probability measure on $(\Omega, \mathcal{A})$, then $(X, \mathcal{A})$ is called a (countably additive, merely finitely additive) *pre-probability space*. If $(X, \mathcal{A})$ is a measurable space and P is a (countably additive, merely finitely additive) probability measure on $(\Omega, \mathcal{A})$, then $(X, \mathcal{A})$ is called a (countably additive, merely finitely additive) *probability space*. An event A in a pre-probability space occurs *almost surely* if $P(A) = 1$.

If $\mathcal{A}$ is an algebra of subsets of X, then a real-valued function $f : X \to \mathbb{R}$ is called $\mathcal{A}$-*measurable* just in case $\{x : f(x) \in B\} \in \mathcal{A}$ for all Borel $B \subseteq \mathbb{R}$. A *random variable* on the pre-measurable space $(X, \mathcal{A})$ is a $\mathcal{A}$-measurable function. If $A \in \mathcal{A}$, then the *indicator function* 1_A for A is defined for all $x \in X$ by

$$1_A(x) = \begin{cases} 0, \text{ if } x \notin A \\ 1, \text{ if } x \in A. \end{cases}$$

Let $(\mathcal{A}_n)_{n\in\mathbb{N}}$ be a filtration of the measurable space $(X, \mathcal{A})$, and let $(f_n)_{n\in\mathbb{N}}$ be a sequence of random variables. We say that the sequence $(f_n)_{n\in\mathbb{N}}$ is *adapted* to $(\mathcal{A}_n)_{n\in\mathbb{N}}$ if f_n is $\mathcal{A}_n$-measurable for all $n \in \mathbb{N}$.

Let f be a random variable on the pre-probability space $(X, \mathcal{A}, P)$. If f takes finitely many values $x_1, ..., x_n$ almost surely, then we say that f is *simple* and define the *expected value* of f to be

$$\mathbb{E}(f) =_{df} \sum_{i=1}^{n} x_i P\big(\{\omega : f(\omega) = x_i\}\big). \tag{3.3}$$

It is a standard result that $\mathbb{E}$ is a well-defined, non-negative, linear functional on the linear space of simple functions. We omit the details.

We note that Markov's inequality, which we have occasion to use below, holds in the finitely additive setting.

Lemma 3.1. *Let $(X, \mathcal{A}, P)$ be a pre-probability space, and let f be a simple, non-negative random variable. Then, for all $r > 0$,*

$$P\big(\{\omega : f(\omega) \geq r\}\big) \leq r^{-1}\mathbb{E}(f). \tag{3.4}$$

Proof. Let $F = \{\omega : f(\omega) \geq r\}$. In general, $f = 1_F f + 1_{F^c} f$, and, since f is non-negative, $f \geq 1_F f \geq 1_F r$, whence $\mathbb{E}(f) \geq P(F)r$. □

Conditional Probability

If $(X, \mathcal{A}, P)$ is a pre-probability space with $A, B \in \mathcal{A}$ and $P(B) > 0$, then the *conditional probability of A given B* is $P(A \mid B) =_{df} P(A \cap B)/P(B)$.

To ensure that conditional probabilities are well-defined in the analysis that follows, we make the following assumption.

Assumption 1. If P is a probability measure on $(\Omega, \mathcal{F})$, $\omega \in \Omega$, and $n \in \mathbb{N}$, then $P(\omega^n) > 0$.

Although this assumption is not without loss of generality, it can be motivated philosophically. If P represents the degrees of belief of a rational agent, then Assumption 1 can be interpreted as an open-mindedness requirement: agents shouldn't assign probability 0 to events whose truth will be settled after no more than n observations. We will discuss this assumption more in the final section.

Fix $A \in \mathcal{A}$, and let P be a probability measure on $(\Omega, \mathcal{F})$. Under Assumption 1, the sequence $(P_n(A))_{n\in\mathbb{N}}$ defined by

$$P_n(A)(\omega) =_{df} P(A \mid \omega^n) \tag{3.5}$$

is a well-defined sequence of random variables on $(\Omega, \mathcal{F})$. Furthermore, this sequence is adapted to the filtration $(\mathcal{C}_n)_{n\in\mathbb{N}}$. In particular, since the cardinality of $\mathcal{C}_n$ is finite, $P_n(A)$ is simple and

$$\mathbb{E}(P_n(A)) = P(A) \tag{3.6}$$

for all $A \in \mathcal{A}$ and $n \in \mathbb{N}$.

3.3.2 Convergence of Random Variables in the Finitely Additive Setting

We begin this subsection by defining the three modes of convergence that will be the focus of our study. After giving the definitions, we discuss the logical relations between them.

Definition 3.1. Let $(X, \mathcal{A}, P)$ be a pre-probability space, and let $(f_n)_{n\in\mathbb{N}}$ be a sequence of random variables. We say that f_n *converges in probability* to the random variable f if for all $\epsilon > 0$,

$$\lim_{n\to\infty} P\big(\{\omega : |f_n(\omega) - f(\omega)| > \epsilon\}\big) = 0. \tag{3.7}$$

We write $f_n \xrightarrow{P} f$.

Definition 3.2. Let $(X, \mathcal{A}, P)$ be a probability space, and let $(f_n)_{n\in\mathbb{N}}$ be a sequence of random variables. We say that f_n *converges almost surely (a.s.)* to the random variable f if,

$$P\big(\{\omega : \lim_{n\to\infty} f_n(\omega) = f(\omega)\}\big) = 1. \tag{3.8}$$

We write $f_n \xrightarrow{a.s.} f$.

Definition 3.3. Let $(X, \mathcal{A}, P)$ be a probability space, and let $(f_n)_{n\in\mathbb{N}}$ be a sequence of random variables. We say that f_n *converges almost uniformly (a.u.)* to the random variable f if for all $\epsilon > 0$ there exists $A_\epsilon \in \mathcal{A}$ such that $P(A_\epsilon) < \epsilon$ and

$$f_n \to f \text{ uniformly on } A_\epsilon^c. \tag{3.9}$$

We write $f_n \xrightarrow{a.u.} f$.

If $(X, \mathcal{A}, P)$ is a countably additive probability space, then it is well-known that

$$f_n \xrightarrow{a.u.} f \iff f_n \xrightarrow{a.s.} f \implies f_n \xrightarrow{P} f. \tag{3.10}$$

Not all of these relations hold for general probability spaces, as we now explain.

The next two lemmas are elementary, but we provide the full proofs in order to call attention to the fact that countable additivity is not used.

Lemma 3.2. *Let $(X, \mathcal{A}, P)$ be a probability space, let $(f_n)_{n\in\mathbb{N}}$ be a sequence of random variables, and let f be a random variable. Suppose that $f_n \xrightarrow{a.u.} f$. Then, for all $\epsilon > 0$*

$$\lim_{n\to\infty} P\big(\big\{\omega : \sup_{i\geq n} |f_i(\omega) - f(\omega)| > \epsilon\big\}\big) = 0. \tag{3.11}$$

Proof. Let $\epsilon > 0$ be given. If $f_n \xrightarrow{a.u.} f$, then there exists $A_\epsilon \in \mathcal{A}$ such that $P(A_\epsilon) < \epsilon$ and $f_n \to f$ uniformly on A_ϵ^c. That is, there is some $N \in \mathbb{N}$ such that $|f_i(\omega) - f(\omega)| \leq \epsilon$ for all $\omega \in A_\epsilon^c$ and all $i \geq N$. Hence, $A_\epsilon^c \subseteq \{\omega : \sup_{i\geq N} |f_i(\omega) - f(\omega)| \leq \epsilon\}$. Thus, if $n \geq N$, then

$$P\big(\big\{\omega : \sup_{i\geq n} |f_i(\omega) - f(\omega)| > \epsilon\big\}\big) \leq P\big(\big\{\omega : \sup_{i\geq N} |f_i(\omega) - f(\omega)| > \epsilon\big\}\big) \leq P(A_\epsilon) < \epsilon,$$

as desired. □

Lemma 3.3. *Let $(X, \mathcal{A}, P)$ be a probability space, let $(f_n)_{n\in\mathbb{N}}$ be a sequence of random variables,*

and let f be a random variable. Then,

$$f_n \xrightarrow{a.u.} f \implies f_n \xrightarrow{a.s.} f \text{ and } f_n \xrightarrow{a.u.} f \implies f_n \xrightarrow{P} f. \tag{3.12}$$

Proof. That $f_n \xrightarrow{a.u.} f \implies f_n \xrightarrow{P} f$ is immediate from Lemma 3.2. For the other implication, assume that $f_n \xrightarrow{a.u.} f$ and for each $n \in \mathbb{N}$ let A_n be such that $P(A_n) < 1/n$ and $f_n \to f$ uniformly on A_n^c. Let $A = \bigcup_{n=1}^{\infty} A_n^c$. Then, $P(A) > 1 - 1/n$ for all n, hence $P(A) = 1$. Moreover, if $\omega \in A$, then $f_n(\omega) \to f(\omega)$, and therefore $f_n \xrightarrow{a.s.} f$. □

The following example shows that the remaining two implications in (3.10) do *not* carry over to the finitely additive setting. That is, in general,

$$\neg\big[f_n \xrightarrow{a.s.} f \implies f_n \xrightarrow{a.u.} f\big] \text{ and } \neg\big[f_n \xrightarrow{a.s.} f \implies f_n \xrightarrow{P} f\big]. \tag{3.13}$$

Example 3.1. *Call a probability space $(\mathbb{N}, 2^{\mathbb{N}}, P)$ a* de Finetti lottery *if $P(A) = 0$ for all $A \subseteq \mathbb{N}$ with finite cardinality. De Finetti lotteries exist and are clearly merely finitely additive.*[7]

For all $i, n \in \mathbb{N}$, let

$$f_n(i) = \begin{cases} 0, & \text{if } i \leq n \\ 1, & \text{otherwise.} \end{cases}$$

If $i \in N$ and $n \geq i$, then $f_n(i) = 0$. Thus, $f_n(i) \to 0$ for all $i \in \mathbb{N}$, and therefore $f_n \xrightarrow{a.s.} 0$.

On the other hand, if $\epsilon > 0$, then $\{i : f_n(i) > \epsilon\} = \{i : i > n\}$ for all n. But

$$P\big(\{i : i > n\}\big) = 1,$$

and therefore f_n does not converge in probability to 0. By Lemma 3.3, f_n does not converge almost uniformly to 0. △

3.3.3 Convergence to the Truth

The following concept is our primary object of study.

[7] See Schirokauer and Kadane (2007).

Definition 3.4. Let P be a probability measure on $(\Omega, \mathcal{F})$. We say that P *converges to the truth* in probability, almost surely, or almost uniformly if, for all $A \in \mathcal{A}$,

$$P_n(A) \xrightarrow{P} 1_A \text{ or } P_n(A) \xrightarrow{a.s.} 1_A \text{ or } P_n(A) \xrightarrow{a.u.} 1_A,$$

respectively.

If P is countably additive, then P converges to the truth in all three modes of convergence. Our aim in the next section is to characterize convergence to the truth for general, finitely additive probability measures.

3.4 Main Results

All of the results reported in this section are for the measurable space $(\Omega, \mathcal{F})$. The first subsection characterizes different modes of convergence to the truth. In the second subsection, we confirm that our characterizing conditions can be satisfied by merely finitely additive probabilities.

3.4.1 Characterizations of Convergence to the Truth

We study the three modes of convergence discussed in 3.3.2 in turn, beginning with convergence in probability.

Convergence in Probability

The next definition is central to the characterization results below.

Definition 3.5. A probability measure P on $(\Omega, \mathcal{F})$ is *approximable by clopen sets*, or *has the approximation property*, if for all $\epsilon > 0$ and $F \in \mathcal{F}$, there exists $C \in \mathcal{C}$ such that $P(F \triangle C) < \epsilon$.

The approximation property can be interpreted as follows. Recall that clopen events—that is, events in $\mathcal{C}$—are those events that can be confirmed or disconfirmed by sampling from Ω finitely many times. They are *verifiable* in the sense that their truth value can be determined by some finite number of observations. Contrast this with an event like $A = \{\omega \in \Omega : \omega(n) = 1 \text{ for infinitely many } n \in \mathbb{N}\}$. For a given $\omega \in \Omega$, there is no n such that by observing the first n coordinates of ω one can determine whether or not $\omega \in A$. In general, it is permissible to assign

probability values to infinitary, non-verifiable events like A. What the approximation property demands is that these values be approximable by values for verifiable events.

The approximation property is both necessary and sufficient for convergence to the truth in probability.

Theorem 3.1. *A probability measure P on $(\Omega, \mathcal{F})$ converges to the truth in probability if and only if P has the approximation property.*

Proof. Suppose that P converges to the truth in probability. Let $A \in \mathcal{F}$ and $\epsilon > 0$ be given. For all $n \in \mathbb{N}$, let $C_n = \{\omega : P(A \mid \omega^n) \geq 1/2\} \in \mathcal{C}_n \subseteq \mathcal{C}$. If $\omega \in A \triangle C_n$, then either $1_A(\omega) = 1$ and $P(A \mid \omega_n) < 1/2$ or $1_A(\omega) = 0$ and $P(A \mid \omega^n) \geq 1/2$. In either case, $|1_A(\omega) - P(A \mid \omega^n)| \geq 1/2$. Thus,

$$P(A \triangle C_n) \leq P\big(\{\omega : |1_A(\omega) - P(A \mid \omega^n)| \geq 1/2|\}\big) \to 0.$$

For some $n \in \mathbb{N}$ and $C_n \in \mathcal{C}$, then, $P(A \triangle C_n) < \epsilon$.

Now suppose that P has the approximation property, and let $A \in \mathcal{F}$ and $\epsilon > 0$ be given. From this it follows that there exists a sequence $(C_m)_{m \in \mathbb{N}}$ of clopen sets such that $P(A \triangle C_m) \to 0$. Now, for all $m, n \in \mathbb{N}$ and $\omega \in \Omega$,

$$|P(A \mid \omega^n) - 1_A(\omega)| \leq P(A \triangle C_m \mid \omega^n) + |P(C_m \mid \omega^n) - 1_{C_m}(\omega)| + |1_{C_m}(\omega) - 1_A(\omega)|.$$

Thus, if $|P(A \mid \omega^n) - 1_A(\omega)| > \epsilon$, then at least one of the quantities $P(A \triangle C_m \mid \omega^n)|$, $|P(C_m \mid \omega^n) - 1_{C_m}(\omega)|$, or $|1_{C_m}(\omega) - 1_A(\omega)|$ is greater than $\epsilon/3$. Hence, using Lemma 3.1 in the final inequality, we get

$$\begin{aligned} P\big(\{\omega : |P(A \mid \omega^n) - 1_A(\omega)| > \epsilon\}\big) &\leq P\big(\{\omega : P(A \triangle C_m \mid \omega^n) > \epsilon/3\}\big) \\ &\quad + P\big(\{\omega : |P(C_m \mid \omega^n) - 1_{C_m}(\omega)| > \epsilon/3\}\big) \\ &\quad + P\big(\{\omega : |1_{C_m}(\omega) - 1_A(\omega)| > \epsilon/3\}\big) \\ &\leq 3\epsilon^{-1} P(A \triangle C_m) \\ &\quad + P\big(\{\omega : |P(C_m \mid \omega^n) - 1_{C_m}(\omega)| > \epsilon/3\}\big) \\ &\quad + P(A \triangle C_m). \end{aligned}$$

By the approximation property, there is an m such that both $3\epsilon^{-1}P(A\triangle C_m)$ and $P(A\triangle C_m)$ are less than $\epsilon/2$. And for all $n \geq m$ and $\omega \in \Omega$, $P(C_m \mid \omega^n) = 1_{C_m}(\omega)$, which implies $P\big(\{\omega : |P(C_m \mid \omega^n) - 1_{C_m}(\omega)| > \epsilon/3\}\big) = 0$. Thus,

$$P\big(\{\omega : |P(A \mid \omega^n) - 1_A(\omega)| > \epsilon\}\big) < \epsilon$$

for all $n \geq m$, so P converges to the truth in probability. □

It is a basic result of measure theory that countably additive probabilities have the approximation property. So, Theorem 3.1 implies that every countably additive probability converges to the truth in probability.

By Theorem 8 of Pomatto et al. (2014), the approximation property is also characteristic of merging of opinions (in probability) in the sense of Blackwell and Dubins (1962). Combining this result with Theorem 3.1, it follows that P converges to the truth in probability if and only if P merges in probability with every probability measure that it dominates. Merging of opinions is discussed a bit more in the final section.

Almost Uniform Convergence

We now turn to almost uniform convergence to the truth. In addition to the approximation property, we will use the following property, which is a limited form of countable additivity.

Definition 3.6. Let P be a probability measure P on $(\Omega, \mathcal{F})$. Let $\delta \in (0,1)$ and $A \in \mathcal{F}$. For all $n \in \mathbb{N}$, let $C_{n,A,\delta} = \{\omega : P(A \mid \omega^n) > \delta\}$, and let $(D_{n,A,\delta})_{n\in\mathbb{N}}$ be the disjoint counterpart of $(C_{n,A,\delta})_{n\in\mathbb{N}}$. If for all $\delta > 0$ and $A \in \mathcal{F}$,

$$P\Big(\bigcup_{n=1}^{\infty} D_{n,A,\delta}\Big) = \sum_{n=1}^{\infty} P\big(D_{n,A,\delta}\big), \tag{3.14}$$

then we say that P is *countably additive on conditional hitting times*, or *has the CHT property.*

By way of motivating the CHT property, it may be helpful to observe that it implies Doob's

maximal inequality. First, note that (3.14) is clearly equivalent to

$$P\Big(\bigcup_{n=1}^{\infty} C_{n,A,\delta}\Big) = \lim_{n\to\infty} P\Big(\bigcup_{i=1}^{n} C_{n,A,\delta}\Big). \tag{3.15}$$

Fix $\delta > 0$, $A \in \mathcal{F}$, and let $X_n(\omega) = P(A \mid \omega^n)$. By a simple calculation, as in Rosenthal (2006, Theorem 14.3.1), one can show that

$$P\big(\{\omega : \max_{1\le i\le n} X_i(\omega) > \delta\}\big) \le P(A)/\delta. \tag{3.16}$$

Then, letting $n \to \infty$ in (3.16) and using (3.15), one obtains the maximal inequality

$$P\big(\{\omega : \sup_{n} X_n(\omega) > \delta\}\big) \le P(A)/\delta. \tag{3.17}$$

Also, note that (3.14) and (3.15) hold trivially for $A \in \mathcal{C}$, since for large enough n, $P(A \mid \omega^n) = 1_A(\omega)$, which implies $C_{n,A,\delta} = A$.

Theorem 3.2. *A probability measure P on $(\Omega, \mathcal{F})$ converges to the truth almost uniformly if and only if P has the approximation property and the CHT property.*

Proof. Suppose that P converges to the truth almost uniformly. By Lemma 3.3 and Theorem 3.1, P has the approximation property. We now show that P also has the CHT property.

To that end, let $\delta \in (0,1)$ and $A \in \mathcal{F}$ be given. Let $D = \bigcup_{n=1}^{\infty} D_{n,A,\delta}$. By finite additivity, $\sum_{n=1}^{\infty} P(A \cap D_{n,A,\delta}) \le P(A \cap D)$. We show the reverse inequality by first calculating, for all $n \in \mathbb{N}$,

$$\begin{aligned} P(A \cap D) &= \sum_{i=1}^{n} P(A \cap D_{i,A,\delta}) + P\Big(A \cap \bigcup_{i=n+1}^{\infty} D_{i,A,\delta}\Big) \\ &\le \sum_{i=1}^{n} P(A \cap D_{i,A,\delta}) + P\big(\{\omega : \sup_{i\ge n} |P(A \mid \omega^i) - 1_A(\omega)| \ge 1-\delta\}\big). \end{aligned} \tag{3.18}$$

The inequality is based on the following reasoning: If $\omega \in A \cap \bigcup_{i=n+1}^{\infty} D_{i,A,\delta}$, then $1_A(\omega) = 1$ and the *first* i for which $P(A \mid \omega_i) > \delta$ is no less than $n+1$, and therefore $P(A \mid \omega^n) \le \delta$, which implies $|P(A \mid \omega^n) - 1_A(\omega)| \ge 1-\delta$. Now, letting $n \to \infty$ in (3.18), using our assumption that P converges

to the truth almost uniformly and Lemma 3.2, we get

$$P(A \cap D) = \sum_{n=1}^{\infty} P(A \cap D_{n,A,\delta}). \tag{3.19}$$

Similarly, for all $n \in \mathbb{N}$, we have

$$\begin{aligned} P(A^c \cap D) &= \sum_{i=1}^{n} P(A^c \cap D_{i,A,\delta}) + P\Big(A^c \cap \bigcup_{i=n+1}^{\infty} D_{i,A,\delta}\Big) \\ &\leq \sum_{i=1}^{n} P(A^c \cap D_{i,A,\delta}) + P\big(\{\omega : \sup_{i \geq n+1} |P(A \mid \omega^i) - 1_A(\omega)| > \delta\}\big). \end{aligned} \tag{3.20}$$

This time the inequality is based on the reasoning: If $\omega \in A^c \cap \bigcup_{i=n+1}^{\infty} D_{i,A,\delta}$, then $1_A(\omega) = 0$ and $P(A \mid \omega^i) > \delta$ *for some* $i \geq n+1$, hence $|P(A \mid \omega^i) - 1_A(\omega)| > \delta$. Letting $n \to \infty$ in (3.20), we get

$$P(A^c \cap D) = \sum_{n=1}^{\infty} P(A^c \cap D_{n,A,\delta}). \tag{3.21}$$

Combining (3.19) and (3.21) yields

$$P(D) = P(A \cap D) + P(A^c \cap D) = \sum_{n=1}^{\infty} \big[P(A^c \cap D_{n,A,\delta}) + P(A \cap D_{n,A,\delta})\big] = \sum_{n=1}^{\infty} P(D_{n,A,\delta}). \tag{3.22}$$

Thus, P has the CHT property.

We now show the converse implication. The proof is a modification of a proof given by Halmos (1950, 49B).[8] Suppose that P has both the approximation and CHT properties. Let $A \in \mathcal{F}$ and $0 < \epsilon, \delta < 1$ be given.

Using the approximation property, let $A_1 \in \mathcal{C}_{n_1}$ be such that $P(A \triangle A_1) < \epsilon\delta/2$. Let $S_1 = A \triangle A_1$, and let $D = \bigcup_{n=1}^{\infty} D_{n,S_1,\delta}$ (Definition 3.6). As $D_{n,S_1,\delta} \in \mathcal{C}_n$, there exist $\omega_1, ..., \omega_{k_n} \in \Omega$ such that the basic clopen sets $\omega_1^n, ..., \omega_{k_n}^n$ are pairwise disjoint and $D_{n,S_1,\delta} = \omega_1^n \cup ... \cup \omega_{k_n}^n$. Then, for all $n \in \mathbb{N}$,

$$P(S_1 \cap D_{n,S_1,\delta}) = \sum_{i=1}^{k_n} P(S_1 \mid \omega_i^n) P(\omega_i^n) > \delta \sum_{i=1}^{k_n} P(\omega_i^n) = \delta P(D_{n,S_1,\delta}). \tag{3.23}$$

[8]This chapter began as a response to Haim Gaifman's conjecture that Bayesian convergence results can be proved in an elementary way, without recourse to martingale theory. I am very grateful to Professor Gaifman for sharing this conjecture with me.

Using (3.23) and the CHT property, we have

$$\begin{aligned}\frac{\epsilon\delta}{2} > P(S_1) \geq P(S_1 \cap D) &= P\Big(S_1 \cap \bigcup_{n=1}^{\infty} D_{n,S_1,\delta}\Big) \\ &\geq \sum_n P(S_1 \cap D_{n,S_1,\delta}) > \delta \sum_n P(D_{n,S_1,\delta}) = \delta P(D). \end{aligned} \tag{3.24}$$

We now let $A_\epsilon = S_1 \cup D$, so that

$$P(A_\epsilon) \leq P(S_1) + P(D) < \frac{\epsilon\delta}{2} + \frac{\epsilon}{2} < \epsilon. \tag{3.25}$$

We aim to show that $P_n(A) \to 1_A$ uniformly on A_ϵ^c. First observe that for all $n \in \mathbb{N}$ and $\omega \in \Omega$

$$|P(A \mid \omega^n) - P(A_1 \mid \omega^n)| \leq P(S_1 \mid \omega^n). \tag{3.26}$$

If $n \geq n_1$, then, because $A_1 \in \mathcal{C}_{n_1}$, (3.26) implies

$$|P(A \mid \omega^n) - 1_{A_1}(\omega)| \leq P(S_1 \mid \omega^n) \tag{3.27}$$

for all $n \in \mathbb{N}$ and $\omega \in \Omega$.

Now let $\omega \in A_\epsilon^c$ be arbitrary. Then, $\omega \in S_1^c$ and $\omega \in D^c$. The first of these implies that $1_A(\omega) = 1_{A_1}(\omega)$, and the second implies that $P(S_1 \mid \omega^n) \leq \delta$ for all $n \in \mathbb{N}$. It follows from (3.27) that if $n \geq n_1$, then

$$|P(A \mid \omega^n) - 1_A(\omega)| \leq \delta.$$

Thus, $P_n(A) \to 1_A$ uniformly on A_ϵ^c, and the theorem is proved. □

Almost Sure Convergence

From the last result, we have the following corollary, which gives conditions that are sufficient for almost sure convergence to the truth.

Corollary 3.1. *If P is a probability measure on $(\Omega, \mathcal{F})$ with the approximation property and the CHT property, then P converges to the truth almost surely.*

Proof. Apply Theorem 3.2 and Lemma 3.3. □

We leave open the problem of finding conditions that are both sufficient and necessary for almost sure convergence to the truth. It might be helpful, however, to note the following easy result: almost sure convergence to the truth can fail only for events with non-extreme probabilities.

Proposition 3.1. *For any probability P on $(\Omega, \mathcal{F})$ satisfying Assumption 1, and any $A \in \mathcal{F}$ with $P(A) \in \{0, 1\}$, $P_n(A) \xrightarrow{a.s.} 1_A$.*

Proof. Suppose $P(A) = 1$. By Assumption 1, $P(A \mid \omega^n) = 1$ for all $\omega \in \Omega$ and $n \in \mathbb{N}$. Thus, for all $\omega \in A$, a P-probability 1 event, we have $P(A \mid \omega^n) \to 1 = 1_A(\omega)$. Similarly, if $P(A) = 0$, then $P(A \mid \omega^n) = 0$ for all $\omega \in \Omega$ and $n \in \mathbb{N}$, and $P(A \mid \omega^n) \to 0 = 1_A(\omega)$ for all $\omega \in A^c$, a P-probability 1 event. □

3.4.2 Existence Results

For all we have shown so far, it could be the case that probability measures with the approximation and CHT properties are countably additive. If that were true, then our work in the previous subsection would not be a real generalization of the countably additive theory. We address this problem in the present subsection by first showing that there are merely finitely additive probabilities that have the approximation property and the CHT property. We then investigate whether these properties are logically independent and show that they are. There exist probabilities with the CHT property but not the approximation property; and there exist probabilities with the approximation property but not the CHT property.

All of the results in this section concern the existence of merely finitely additive probabilities on the measurable space $(\Omega, \mathcal{F})$. The existence of these objects is known to be independent of the Zermelo-Fraenkel axioms of set theory without the Axiom of Choice, so the results to follow are highly non-constructive.[9] One of our main tools is a result due to Plachky (1976), which we now introduce.

Let $\mathcal{A}$ be an algebra of subsets of X, and let $\mathcal{B}$ be a sub-algebra of $\mathcal{A}$. If P is a probability measure on $\mathcal{A}$, then we let $P_{\mathcal{B}}$ denote the restriction of P to $\mathcal{B}$. If Q is a probability measure on $\mathcal{B}$, then P is an *extension* of Q from $\mathcal{B}$ to $\mathcal{A}$ if P is a probability measure on $(X, \mathcal{A})$ and $P_{\mathcal{B}} = Q$. The set of extensions of Q from $\mathcal{B}$ to $\mathcal{A}$ is denoted by $E(Q, \mathcal{B}, \mathcal{A})$. Note that this set is convex

[9] See Schechter (1996, 29.37).

and, by the Krein-Milman theorem, has extreme points. The set of extreme points of $E(Q, \mathcal{B}, \mathcal{A})$ is denoted by $exE(Q, \mathcal{B}, \mathcal{A})$.

Lemma 3.4 (Plachky 1976, Theorem 1)**.** *Let $\mathcal{A}$ and $\mathcal{B}$ be algebras of subsets of X with $\mathcal{B} \subseteq \mathcal{A}$, and let Q be a probability measure on $(X, \mathcal{B})$. Let $P \in E(Q, \mathcal{B}, \mathcal{A})$. Then, $P \in exE(Q, \mathcal{B}, \mathcal{A})$ if and only if for all $A \in \mathcal{A}$ and $\epsilon > 0$ there is some $B \in \mathcal{B}$ such that $P(A \triangle B) < \epsilon$.*

Note that an immediate corollary of Lemma 3.4 is that a probability measure P on $(\Omega, \mathcal{F})$ has the approximation property if and only if $P \in ex(P_{\mathcal{C}}, \mathcal{C}, \mathcal{F})$.

Finally, before showing the first result, we should note that a related result, whose proof also uses Plachky's theorem, appears in Pomatto et al. (2014) (Theorem 10). An immediate corollary of that result and our Theorem 3.1 is that there exist uncountably many merely finitely additive probabilities that converge to the truth in probability. The result that follows improves on this.

Theorem 3.3. *There are uncountably many merely finitely additive probabilities on $(\Omega, \mathcal{F})$ that satisfy Assumption 1 and converge to the truth almost uniformly.*

Proof. Let Q be the countably additive probability measure on $(\Omega, \mathcal{F})$ defined by $Q(\omega^n) = 2^{-n}$ for all $\omega \in \Omega$ and $n \in \mathbb{N}$, so that Q satisfies Assumption 1. Let $\mathcal{A}$ denote the algebra of subsets of Ω generated by the open sets. Note that $\mathcal{C} \subset \mathcal{A}$. We proceed by showing a number of claims.

Claim (a). Every member of $E(Q_{\mathcal{A}}, \mathcal{A}, \mathcal{F})$ satisfies Assumption 1.

This is immediate from the definitions.

Claim (b). $|exE(Q_{\mathcal{A}}, \mathcal{A}, \mathcal{F})| \geq 2^{\aleph_0}$.

By Theorem 1 of Lipecki (2001), $|exE(Q_{\mathcal{A}}, \mathcal{A}, \mathcal{F})| = \mathfrak{n}^{\aleph_0}$ for some cardinal $\mathfrak{n}$. To establish the claim, then, it suffices to show that $|exE(Q_{\mathcal{A}}, \mathcal{A}, \mathcal{F})| \geq 2$. Start by viewing our probabilities as forming a compact subset of the topological vector space of signed, finitely additive measures on $(\Omega, \mathcal{F})$ that are bounded in the variation norm (this space can, in turn, be viewed as the dual space of the space of bounded, measurable functions on $(\Omega, \mathcal{F})$). If $|exE(Q_{\mathcal{A}}, \mathcal{A}, \mathcal{F})| = 1$, then the Krein-Milman theorem would imply that $|E(Q_{\mathcal{A}}, \mathcal{A}, \mathcal{F})| = 1$, since $E(Q_{\mathcal{A}}, \mathcal{A}, \mathcal{F})$ is the closed convex hull of its extreme points. Thus, to finish proving the claim, it now suffices to exhibit a member of $E(Q_{\mathcal{A}}, \mathcal{A}, \mathcal{F})$ that is not Q. To that end, let D be a countable dense subset of Ω. Then, $D \notin \mathcal{A}$. Let $\mathcal{D}$ be the algebra generated by $\mathcal{A} \cup \{D\}$. Every member of $\mathcal{D}$ is of the form $(A_1 \cap D) \cup (A_2 \cap D^c)$,

$A_1, A_2 \in \mathcal{A}$. Following Pomatto et al. (2014, Theorem 10), define P on $(\Omega, \mathcal{D})$ by

$$P\big((A_1 \cap D) \cup (A_2 \cap D^c)\big) = Q_{\mathcal{A}}(A_1).$$

Then, P is a well-defined finitely additive extension of $Q_{\mathcal{A}}$ to $\mathcal{D}$. But note that $P(D) = 1$, whereas $Q(D) = 0$. So, by taking any extension of P from $\mathcal{D}$ to $\mathcal{F}$, we have a finitely additive extension of $Q_{\mathcal{A}}$ from $\mathcal{A}$ to $\mathcal{F}$ that is not identical to Q, and the claim is proved.

Claim (c). If $P \in exE(Q_{\mathcal{A}}, \mathcal{A}, \mathcal{F})$, then for all $F \in \mathcal{F}$ and $\epsilon > 0$, there exists $A \in \mathcal{A}$ such that $P(F \triangle A) < \epsilon$.

Apply Lemma 3.4.

Claim (d). If $P \in exE(Q_{\mathcal{A}}, \mathcal{A}, \mathcal{F})$, then P has the approximation property.

Let $F \in \mathcal{F}$ and $\epsilon > 0$ be given. Using claim (c), there exists $A \in \mathcal{A}$ such that $P(F \triangle A) < \epsilon/2$. As Q is countably additive, there exists $C \in \mathcal{C}$ such that $Q(A \triangle C) < \epsilon/2$. But $A \triangle C \in \mathcal{A}$, so, since $P \in E(Q_{\mathcal{A}}, \mathcal{A}, \mathcal{F})$, we have $P(A \triangle C) = Q(A \triangle C)$. Thus,

$$P(F \triangle C) \leq P(F \triangle A) + P(A \triangle C) < \epsilon,$$

which proves the claim.

Claim (e). If $P \in exE(Q_{\mathcal{A}}, \mathcal{A}, \mathcal{F})$, then P has the CHT property.

In the definition of the CHT property (Definition 3.6), the sequence $(D_{n,A,\delta})_{n \in \mathbb{N}}$ and the set $\bigcup_{n=1}^{\infty} D_{n,A,\delta}$ are contained in $\mathcal{A}$. Since Q is countably additive and $P \in E(Q_{\mathcal{A}}, \mathcal{A}, \mathcal{F})$, P is countably additive on $\mathcal{A}$. Thus, P has the CHT property.

Claim (f). At most one probability measure in $exE(Q_{\mathcal{A}}, \mathcal{A}, \mathcal{F})$ is countably additive.

Since $\sigma(\mathcal{A}) = \mathcal{F}$, any countably additive probability that agrees with Q on $\mathcal{A}$ must also agree with Q on $\mathcal{F}$.

Now, by Theorem 3.2 and claims (a), (d) and (e), every $P \in exE(Q_{\mathcal{A}}, \mathcal{A}, \mathcal{F})$ satisfies Assumption 1 and converges to the truth almost uniformly. By claims (b) and (f), uncountably many $P \in exE(Q_{\mathcal{A}}, \mathcal{A}, \mathcal{F})$ are merely finitely additive. □

It follows from Lemma 3.3 that there are uncountably many merely finitely additive probabilities that satisfy Assumption 1 and converge to the truth almost surely (resp. in probability).

The next task is to determine whether the CHT property and the approximation property are logically independent. The next result shows that the CHT property does not imply the approximation property.

Theorem 3.4. *There are uncountably many probability measures on* $(\Omega, \mathcal{F})$ *that satisfy Assumption 1 and have the CHT property but do not have the approximation property.*

Proof. We use the same notation that was used in the proof of Theorem 3.3. If $P \in E(Q_{\mathcal{A}}, \mathcal{A}, \mathcal{F})$, then P satisfies Assumption 1. Moreover, because P is countably additive on $\mathcal{A}$, P has the CHT property. In the proof of Theorem 3.3, we showed that there are at least two extreme points of $E(Q_{\mathcal{A}}, \mathcal{A}, \mathcal{F})$. If P is a non-trivial convex combination of these—and there are uncountably many such P—then P is not an extreme point. Then, by Lemma 3.4, there are $F \in \mathcal{F}$ and $\epsilon > 0$ such that $P(F \triangle A) \geq \epsilon$ for all $A \in \mathcal{A}$. As $\mathcal{C} \subset \mathcal{A}$, $P(F \triangle C) \geq \epsilon$ for all $C \in \mathcal{C}$. So P does not have the approximation property. □

We conclude this section by showing the converse of Theorem 3.4. The approximation proprety does not imply the CHT property. In fact, our result is a bit stronger than this. We will show that the approximation property does not imply almost sure convergence to the truth.

Theorem 3.5. *There are uncountably many probability measures on* $(\Omega, \mathcal{F})$ *that satisfy Assumption 1 and have the approximation property but do not converge to the truth almost surely, and therefore do not have the CHT property.*

Proof. The second claim follows from the first, by Theorem 3.2 and Lemma 3.3, so it remains only to prove the first claim.

Let Q be defined as in the proof of Theorem 3.3, i.e. $Q(\omega^n) = 2^{-n}$ for all $\omega \in \Omega$ and $n \in \mathbb{N}$. Let D be an infinite discrete topological subspace of Ω such that $Q(D \cup L_D) = 0$, where L_D is the closed set of limit points of D. (For example, we can take D to contain all and only those ω_n defined by $w_n(i) = 1_{\{n\}}(i)$, $i \in \mathbb{N}$, and then L_D contains just the sequence that is constantly 0.) Note that since D is discrete, $D \subseteq L_D^c$. Let $\mathcal{U}$ be a non-principal ultrafilter containing D, and let u be the extreme valued probability induced by $\mathcal{U}$, i.e. $u(F) = 1_{\mathcal{U}}(F)$ for all $F \in \mathcal{F}$. In particular, $u(D) = 1$. If $\alpha \in (0, 1)$, let $P_\alpha = \alpha u + (1 - \alpha)Q$. Then, P_α satisfies Assumption 1 because Q does.

We will show that, for all $\alpha \in (0,1)$, P_α has the approximation property but does not converge to the truth almost surely. Let $\alpha \in (0,1)$ be given.

Claim (a). If $C \in \mathcal{C}$ and $C \subseteq L_D^c$, then $u(C) = 0$.

Since C is compact and $C \cap L_D = \emptyset$, the intersection $C \cap D$ is finite. (If $C \cap D$ were infinite, it would contain one of its limit points because it is compact; but then D would contain one of its limit points, contradicting $D \subseteq L_D^c$.) Since $\mathcal{U}$ is non-principal, it follows that $C \cap D \notin \mathcal{U}$, and therefore, since $D \in \mathcal{U}$, $C \notin \mathcal{U}$. Thus, $u(C) = 0$.

Claim (b). P_α has the approximation property.[10]

Let $F \in \mathcal{F}$ and $\epsilon > 0$ be given. There are two cases to consider. First, suppose that $u(F) = 0$. Using the fact that Q is a regular Borel measure and $Q(L_D) = 0$, we can find a compact set $K \subseteq F - L_D$ such that $Q(K) > Q(F) - \epsilon/2$. Then, using the regularity of Q once more and the compactness of K, we can find a clopen set C such that $K \subseteq C \subseteq L_D^c$ and $Q(C) < Q(K) + \epsilon/2$. Then,

$$Q(F \triangle C) \leq Q(F \triangle K) + Q(K \triangle C) < \epsilon.$$

By claim (a), $u(C) = 0$. Thus,

$$P_\alpha(F \triangle C) = \alpha u(F \triangle C) + (1-\alpha) Q(F \triangle C) < u(F) + \epsilon = \epsilon.$$

Second, suppose that $u(F) = 1$. Using the regularity of Q, the fact that $Q(L_D) = 0$, and the compactness of L_D, we can find a compact set K such that $L_D \subseteq K \subseteq F$ and $Q(K) > Q(F) - \epsilon/2$. Then, using the regularity of Q again and the compactness of K, we can find a clopen set $C \supseteq K$ such that $Q(C) < Q(K) + \epsilon/2$. It follows that $Q(F \triangle C) < \epsilon$. Since $C \supseteq L_D$ implies $C^c \subseteq L_D^c$, we have, by claim (a), $u(C^c) = 0$. So, $u(C) = 1$, and therefore $u(C \cap F) = 1$. Then,

$$P_\alpha(F \triangle C) = \alpha u(F \triangle C) + (1-\alpha) Q(F \triangle C) = (1-\alpha) Q(F \triangle C) < \epsilon.$$

This completes the proof of the claim.

Claim (c). If $\omega \in D$, then $P_\alpha(D \mid \omega^n) \to 0$.

[10] I am grateful to Taras Banakh for suggesting the construction of P_α to me in a slightly different context, and for help with the argument establishing claim (b). Any errors that appear are mine alone. For a survey of the properties of regular Borel measures, see Aliprantis and Border (2006, Chapter 12).

If $\omega \in D$, then, since D is discrete, there is some open subset G of Ω such that $D \cap G = \{\omega\}$. Express G as a countable, pairwise disjoint union of basic clopen sets: $G = \omega_1^{m_1} \cup \omega_2^{m_2} \cup ....$ Since $\omega \in G$, $\omega \in \omega_i^{m_i}$ for exactly one $i \in \mathbb{N}$. If $n \geq m_i$, then $D \cap \omega^n \subseteq D \cap \omega_i^{m_i} \subseteq D \cap G$, whence $D \cap \omega^n = \{\omega\}$. Now, for all $n \geq m_i$,

$$P_\alpha(D \mid \omega^n) = \frac{\alpha u(D \cap \omega^n) + (1-\alpha)Q(D \cap \omega^n))}{\alpha u(\omega^n) + (1-\alpha)Q(\omega^n)} = \frac{\alpha u(\{\omega\}) + (1-\alpha)Q(\{\omega\})}{\alpha u(\omega^n) + (1-\alpha)2^{-n}} = 0,$$

because $u(\{\omega\}) = 0 = Q(\{\omega\})$. This establishes the claim.

To conclude the proof of the theorem, observe that $P_\alpha(D) = \alpha > 0$. Thus, by claim (c), the collection of $\omega \in \Omega$ for which $P_\alpha(D \mid \omega^n)$ does not converge to $1_D(\omega)$ contains D, and therefore has positive measure according to P_α. □

To sum up, we have shown that there are uncountably many merely finitely additive probabilities that converge to the truth almost uniformly, and therefore almost surely and in probability. And we have shown that the properties we used to characterize different modes of convergence to the truth—the approximation property and the CHT property—are logically independent, with uncountably many witnesses to this fact in both directions.

3.5 Questions for Future Research

There are a number of open questions that need to be addressed by future research. We have left open the problem of characterizing almost sure convergence to the truth, though Corollary 3.1 provides sufficient conditions. Solving this problem would complete the theory of convergence to the truth without countable additivity that we have developed in this paper. Having completed this project, it would then be natural to study finitely additive merging of opinions results under the various modes of convergence included in this paper. As mentioned above, the problem of characterizing merging of opinions in probability has already been settled by Pomatto et al. (2014). The characterizing condition is the same as that for convergence to the truth in probability, namely the approximation property. This suggests investigating whether the conjunction of the approximation property and the CHT property also characterizes almost uniform merging of opinions. There is also the matter of generalizing the results above by considering more general (pre-)measurable spaces

and relaxing Assumption 1. To do this, one needs a general theory of finitely additive conditional probability. Perhaps full conditional probabilities in the sense of Dubins (1975) can be applied to this end. In short, there are plenty of interesting open problems in this area for researchers to look forward to.

Chapter 4

Speed-Optimal Induction and Dynamic Coherence

4.1 Introduction

Reichenbach held that the inductive methods used in the sciences are essentially rules for estimating probabilities. Probabilities, in turn, received a frequency interpretation, and this led Reichenbach to regard the discovery of limiting relative frequencies as a primary aim of scientific inquiry.[1]

Reichenbach advocated a particularly simple inductive method for predicting frequencies, the *straight rule*. Using the straight rule, he attempted a "pragmatic vindication" of induction in response to Humean skepticism.[2] We must concede to Hume, Reichenbach thought, that we cannot be certain of nature's regularity. But if nature is regular, then the straight rule will reveal this to us in the limit of inquiry. In particular, if the relative frequency of an outcome in a repeated experiment approaches a stable limit, then the straight rule's conjectures about the outcome's frequency necessarily approach the same limit.

The chief problem with Reichenbach's account of inductive success—convergence to the correct limiting relative frequencies, when the limits exist—is that it places no constraints whatsoever on the kinds of inductive inferences that can be made in the short-term.[3] Arbitrary conjectures in response to a finite amount of data can always be extended in a way that secures convergence in the long run.

In view of this limitation of Reichenbach's account, it is natural to ask whether stronger criteria of inductive success are able to induce significant short-term constraints. This line of thought is

[1] Reichenbach (1938, 1949). Also see van Fraassen (2000).

[2] Salmon (1991).

[3] Salmon (1966). For a more contemporary take on some of the limitations of Reichenbach's account, see Huttegger (2017a, 3.1).

pursued by Juhl (1994), who introduces the notion of *speed-optimal convergence*. Juhl shows that the straight rule is speed-optimal in his sense and that there are inductive methods that, although convergent, are not speed-optimal. Considered as a criterion of inductive success, then, speed-optimality places more constraints on inductive methodology than the Reichenbachian account, which requires mere convergence.

Importantly, however, Juhl's analysis leaves open one of the questions that motivates it. Namely, does requiring speed-optimal convergence of one's inductive method induce significant short-term constraints? Or can arbitrary short-term behavior always be extended in a speed-optimal way? The primary aim of this paper is to answer these questions. Speed-optimal convergence *does* give rise to short-term constraints; not all short-term behavior can be extended without loss of speed-optimality. Rather surprisingly, the short-term constraint that can be derived from the requirement of speed-optimal convergence is one of great independent interest in the philosophy of induction: *dynamic coherence*. Along the way to proving this result, we will answer another open question of Juhl's by providing a complete characterization of the speed-optimal inductive methods.[4]

We begin, in the next section, by presenting the standard mathematical framework for making our introductory remarks precise. In section 4.3, we discuss speed-optimality and present our characterization result. Section 4.4 contains our main result: short-term inductive behavior can be extended in a speed-optimal way if and only if it is dynamically coherent. We conclude, in section 4.5, by discussing some connections to probabilistic learning and martingales. Proofs are in the Appendix.

4.2 Mathematical Preliminaries

Let $\mathcal{C}$ be the collection of all binary sequences. Sequences in $\mathcal{C}$ will be denoted by variants of $\sigma = (\sigma_1, \sigma_2, ...)$.

Let $\mathcal{S} = \bigcup_n \{0,1\}^n$ be the collection of all finite length binary sequences, which we will call *strings* from now on. Generic elements of $\mathcal{S}$ will be denoted by variants of $s = (s_1, ..., s_n)$. Let $|s|$ denote the length of the string s.

[4]We should note at the outset that although formal learning theory has made many advances in the years since Juhl's paper, including in the study of fast and efficient inductive methods (Kelly, 1996; Schulte, 1999a,b), the results in that literature do not, to the best of our knowledge, provide immediate and fully satisfactory answers to the questions that Juhl raises about learning limiting relative frequencies.

If $\sigma \in \mathcal{C}$ and $n \in \mathbb{N}$, let $\sigma^n = (\sigma_1, ..., \sigma_n) \in \mathcal{S}$ denote the initial segment of σ of length n. Similarly, if $s \in \mathcal{S}$ and $n \leq |s|$, then $s^n = (s_1, ..., s_n)$ is the initial segment of s of length n. If $s \in \mathcal{S}$ and $a \in \{0, 1\}$, let $sa = (s_1, ..., s_{|s|}, a)$. A sequence σ (resp. string s') *extends* a string s if $\sigma^{|s|} = s$ (resp. $s'^{|s|} = s$).

Let $\mathcal{C}_{\text{conv}}$ denote the collection of sequences in $\mathcal{C}$ such that the limiting relative frequency of 1s exists. That is, if $\sigma \in \mathcal{C}_{\text{conv}}$, then

$$\ell(\sigma) := \lim_{n \to \infty} \frac{\sigma_1 + ... + \sigma_n}{n}$$

exists.

An *inductive method* ϕ is a function of $\mathcal{S}$ into $[0, 1]$. The value $\phi(s)$ is interpreted as a conjecture about the limiting relative frequency of 1s based on an observation of s. Note that inductive methods are just arbitrary functions from strings into the unit interval. In particular, they are not assumed to have any probabilistic structure.

An inductive method is called *convergent* if

$$\lim_{n \to \infty} \phi(\sigma^n) = \ell(\sigma)$$

for all $\sigma \in \mathcal{C}_{\text{conv}}$. The conjectures of convergent methods approach actual limiting relative frequencies whenever the limits exist.

The *straight rule sr* is the inductive method defined by

$$sr(s) = \frac{s_1 + ... + s_{|s|}}{|s|}$$

for all $s \in \mathcal{S}$. The straight rule always conjectures the observed relative frequency of 1s. It is immediate from the definitions of the relevant terms that the straight rule is convergent. Conjectures equal to observed frequencies are guaranteed to converge to actual limiting relative frequencies whenever the limits exist. It is in this sense that the straight rule is supposed to vindicate induction: we cannot be certain that regularities will emerge as observations unfold—for all we know, the relative frequency of 1s may oscillate forever—but if there is regularity, then the straight rule is sure to identify it in the limit.

The claim that convergence does not constrain short-term behavior can be demonstrated as follows. Let $s_1, ..., s_n$ be an arbitrary, finite collection of strings, and let $r_1, ..., r_n$ be an arbitrary collection of real numbers in $[0, 1]$. Then, there is a convergent inductive method ϕ such that $\phi(s_i) = r_i$ for all $i \in \{1, ..., n\}$. For example, let ϕ agree with sr on all strings besides, perhaps, $s_1, ..., s_n$. Since conjectures on $s_1, ..., s_n$ are irrelevant to ϕ's behavior in the limit, convergence is consistent with arbitrarily divergent predictions in the short-term.

4.3 Speed-Optimality

Pursuing a remark from Salmon (1966), Juhl (1994) raises the question whether we can derive short-term constraints by requiring more of our inductive methods than mere convergence. If, following Reichenbach, a primary aim of scientific inquiry is to learn frequencies, then it seems reasonable to favor those inductive methods that converge as quickly as possible. One can make this idea precise as follows.

For all $\sigma \in \mathcal{C}_{\text{conv}}$, $\epsilon > 0$, and convergent inductive methods ϕ, ϕ', we write $\phi \succ_{\sigma,\epsilon} \phi'$ if and only if there exists $m \in \mathbb{N}$ such that

$$|\phi(\sigma^n) - \ell(\sigma)| \leq \epsilon < |\phi'(\sigma^m) - \ell(\sigma)| \text{ whenever } n \geq m.$$

In other words, $\phi \succ_{\sigma,\epsilon} \phi'$ holds if and only if ϕ is within, and forever remains within, ϵ of $\ell(\sigma)$ strictly before ϕ'. When $\phi \succ_{\sigma,\epsilon} \phi'$ holds, we sometimes say that ϕ *beats* ϕ' *on* σ, ϵ.

We say that an inductive method ϕ is *faster* than another inductive method ϕ' if and only if

$$\exists \sigma \in \mathcal{C}_{\text{conv}}, \epsilon > 0 : \ \phi \succ_{\sigma,\epsilon} \phi' \text{ and } \forall \sigma \in \mathcal{C}_{\text{conv}}, \epsilon > 0 : \neg[\phi' \succ_{\sigma,\epsilon} \phi].$$

If ϕ is faster than ϕ', then ϕ beats ϕ' on some σ, ϵ and ϕ' does not beat ϕ on any σ, ϵ.

We say that a convergent inductive method ϕ is *speed-optimal* if and only if there does not exist a convergent inductive method ϕ' that is faster than ϕ. An inductive method is speed-optimal in this sense if there is no way of modifying its conjectures that leads to faster convergence.

We now summarize the known facts about speed-optimal convergence and provide an additional

example of the property. An inductive method ϕ is called *monotonic* if and only if

$$\phi(s1) \geq \phi(s) \quad \text{and} \quad \phi(s0) \leq \phi(s)$$

for all $s \in \mathcal{S}$. In other words, monotonic methods do not decrease (resp. increase) their conjectures about the frequency of 1s in response to observing an additional 1 (resp. 0).

Facts (Juhl, 1994). If an inductive method is convergent and monotonic, then it is speed-optimal. Hence, the straight rule is speed-optimal. There exist non-monotonic, speed-optimal inductive methods. There exist convergent inductive methods that are not speed-optimal.

Example. Say that an inductive method ϕ is *Laplacian* if there exist parameters $\alpha_1, \alpha_2 \geq 0$ such that for all $s \in \mathcal{S}$

$$\phi(s) = \frac{\sum_i s_i + \alpha_1}{|s| + \alpha_1 + \alpha_2}.$$[5]

The straight rule is Laplacian with parameters $\alpha_1 = \alpha_2 = 0$. Intuitively, Laplacian methods are biased straight rules, with the biases encoded by the parameters α_1 and α_2. It is a straightforward exercise to show that Laplacian methods are convergent and monotonic, and therefore speed-optimal.

In view of the partial results recorded in the Facts above, Juhl asks, "Exactly which [convergent inductive methods] are speed-optimal?" (862). In the remainder of this section, we provide an answer to this question. In the next section, we extend our answer to show that speed-optimal convergence gives rise to short-term constraints.

The formal definition of our characterizing condition for speed-optimality is somewhat technical, but the idea behind it is simple and easy to explain: the conjectures of speed-optimal methods are *rigid* in the sense that they cannot be changed without sacrificing speed. Intuitively, our characterization result shows that if ϕ's conjecture at s can be changed without the resulting method being any slower than ϕ, then ϕ cannot have been speed-optimal in the first place; and, conversely, if any change to ϕ's conjecture at s results in a slower inductive method, then ϕ is speed-optimal.

[5]This formula generalizes Laplace's rule of succession and was developed independently by Johnson (1924; 1932) and Carnap (1950; 1952). See Huttegger (2017a) for more details.

We'll now introduce the formal definition of rigidity. If we are given an inductive method ϕ, a string $s \in \mathcal{S}$, and a sequence $\sigma \in \mathcal{C}_{conv}$ extending s, we write

$$\epsilon_{\phi,s,\sigma} = \sup_{n \geq |s|} |\phi(\sigma^n) - \ell(\sigma)|,$$

which is the largest distance between $\phi(\sigma^n)$ and $\ell(\sigma)$ after time $|s|$. We also define

$$I_{\phi,s,\sigma} = [\ell(\sigma) - \epsilon_{\phi,s,\sigma}, \ell(\sigma) + \epsilon_{\phi,s,\sigma}] \cap [0,1],$$

which is the smallest closed interval, centered at $\ell(\sigma)$, that contains $\phi(\sigma^n)$ for all $n \geq |s|$. Finally, we define

$$I_{\phi,s} = \bigcap_{\sigma \in \mathcal{C}_{conv}} I_{\phi,s,\sigma}.$$

One important thing to note is that, by the definitions just given, $\phi(s) \in I_{\phi,s}$ for all ϕ and s. In particular, $I_{\phi,s}$ is nonempty.

Intuitively, the closed interval $I_{\phi,s}$ represents ways that ϕ's conjecture at s can be changed without sacrificing speed. To see this, suppose that $x \in I_{\phi,s}$ and define a new method ϕ' from ϕ by $\phi'(s) = x$ and $\phi'(t) = \phi(t)$ for all $t \neq s$. The method ϕ' is the result of setting ϕ's conjecture at s to x and leaving ϕ's other conjectures unchanged. By unpacking the definitions above, we can see that ϕ is not faster than ϕ'. In particular, ϕ does not beat ϕ' on any σ, ϵ. To spell this out in a bit more detail, consider any sequence $\sigma \in \mathcal{C}_{\text{conv}}$ that extends s (ϕ cannot beat ϕ' on σ, ϵ if σ is *not* an extension of s because ϕ and ϕ' make the same conjectures about all initial segments of such a σ, by construction). Now, $\phi'(s)$ is in the interval $I_{\phi,s,\sigma}$ by definition. The radius of the interval $I_{\phi,s,\sigma}$ is (by definition) the *smallest* ϵ such that $\phi(\sigma^n)$ is within ϵ of $\ell(\sigma)$ at all times n after $|s|$. Since $\phi'(s)$ lies within this interval, we do *not* have

$$|\phi(\sigma^n) - \ell(\sigma)| \leq \epsilon < |\phi'(s) - \ell(\sigma)|, \ \forall n \geq |s|$$

for any ϵ. So, since s is the only string on which the conjectures of ϕ and ϕ' differ, ϕ is not faster than ϕ'. Thus, the numbers $x \neq \phi(s)$ in the interval $I_{\phi,s}$ are *speed-preserving alternatives* to $\phi(s)$ in the sense that changing ϕ's conjecture at s to x does not result in a slower inductive method.

Rigid inductive methods do not have speed-preserving alternatives. Formally, we say that an inductive method ϕ is *rigid* if $I_{\phi,s} = \{\phi(s)\}$ for all $s \in \mathcal{S}$, i.e. $I_{\phi,s}$ is degenerate. Rigidity characterizes speed-optimality.

Theorem 4.1. *A convergent inductive method is speed-optimal if and only if it is rigid.*

We will explore rigidity further in the next section, but for now we turn to another question that Juhl raises.

4.4 Dynamic Coherence

The question that motivated us at the outset is whether requiring one's inductive method to be speed-optimally convergent places any significant constraints on the method's predictions in the short-term. Juhl (1994) asks precisely this as an open question at the end of his paper:

> Is *any* short-term behavior compatible with speed-optimality?...If a negative answer to this question can be proved, then we will have established the existence of short-run norms on estimation rules. If non-trivial short-term norms can be shown to be induced by the requirement of speed-optimality, then the chief intuitive objection to Reichenbach's attempts to 'vindicate induction' would be answered (862).

The aim of this section is to show that, indeed, a negative answer to Juhl's question can be established. Speed-optimality does induce a non-trivial and, we will argue, particularly interesting short-term inductive constraint.

Let us begin by formalizing the problem. Call a function f from a finite subset A of $\mathcal{S}$ into $[0,1]$ a *partial inductive method*. Partial inductive methods represent prediction behavior in the short-term. Since a partial inductive method f is defined on a finite set, there is some long-run time horizon n such that f is undefined for all strings of length more than n. An inductive method $\phi : \mathcal{S} \to [0,1]$ is an *extension* of $f : A \to [0,1]$ if $\phi(s) = f(s)$ for all $s \in A$, and we say that ϕ *extends* f. Now, the question whether any short-term behavior is compatible with speed-optimality becomes: *Is every partial inductive method extended by some speed-optimal convergent inductive method?* If the answer to this question is negative, and it can be shown that exactly those partial inductive methods with property P admit speed-optimal extensions, then we say that

speed-optimality induces the short-term constraint P. We will now start working towards a result along these lines.

Our short-term constraint turns out to be closely related to a central idea in the philosophy of induction, and in theorizing about rational learning more generally. The idea is that the conjectures that a rational inductive method makes at a given time are constrained in a particular way by the conjectures that it might make in the future. It is irrational, the idea goes, to conjecture x now while at the same time expecting to conjecture $y \neq x$ in the future *no matter what new data one observes between now and then.*

This general idea has been formalized in a number of ways across several disciplines. In the philosophy of probability and formal epistemology, the *principle of reflection* captures the idea. It says that a rational agent's conditional probability for A, given that her probability for A will be x after learning some new evidence, must be equal to x.[6] In the theory of finitely additive probability, researchers have been puzzled by violations of *conglomerability*: an unconditional, finitely additive probability value need not reside in the interval spanned by its conditional probability values, given members of a countably infinite partition.[7] In decision theory, Savage's *sure-thing principle* says that if option 1 is preferred to option 2 conditional on every member of some partition, then option 1 ought to be preferred to option 2 unconditionally.[8] Similar principles of *dynamic consistency* appear in the economics literature.[9] In statistics, results due to Lane and Sudderth (1984; 1985) show that probability estimates are *dynamically coherent* (avoid Dutch book) just in case they are contained within the closed, convex hull of possible future estimates.

The condition that we articulate is similar to the result just mentioned, and we borrow the terminology accordingly. Let $\mathcal{S}_n$ denote the collection of strings of length at most n. We say that a partial inductive method $f : \mathcal{S}_n \to [0, 1]$ is *dynamically coherent* (or sometimes simply *coherent*) if for all $s \in \mathcal{S}_{n-1}$

$$f(s0) \leq f(s) \leq f(s1) \text{ or } f(s1) \leq f(s) \leq f(s0).$$

[6] van Fraassen (1984, 1999); van Fraassen and Halpern (2016); Huttegger (2013, 2014).

[7] de Finetti (1972); Dubins (1975); Schervish et al. (1984); Kadane et al. (1996). A related phenomenon in the theory of imprecise probability is *dilation* (Seidenfeld and Wasserman, 1993; Pedersen and Wheeler, 2014, 2015).

[8] Savage (1972). Gaifman (2013) discusses connections between some of the phenomena mentioned above. Also see Gaifman and Vasudevan (2012).

[9] Epstein and Le Breton (1993); Epstein and Schneider (2003).

For example, dynamic coherence rules out the possibility of conjecturing 0.5 now and 0.6 after the next observation no matter what is observed. Put another way, the conjectures of dynamically coherent methods are always contained within the interval spanned by the conjectures that might be made after observing more data. We note that monotonic inductive methods are coherent.

We now have the following preliminary result.

Lemma 4.1. *Let $n \in \mathbb{N}$ and $f : \mathcal{S}_n \to [0,1]$. Then there exists a speed-optimal convergent inductive method that extends f if and only if f is dynamically coherent.*

Lemma 4.1 provides a partial answer to the question that we raised above. There are partial inductive methods that do not have speed-optimal extensions. In particular, any method that fails to be dynamically coherent cannot be extended without sacrificing speed-optimality. We see that speed-optimality, then, induces dynamic coherence in the short-term *for all partial inductive methods with domains of the form $\mathcal{S}_n$*.

Removing this last proviso, so that the conclusion of Lemma 4.1 applies to *all* partial inductive methods, requires a modest generalization of the definition of dynamic coherence. Given $s, t \in \mathcal{S}$, we write $s \sqsubseteq t$ (resp. $s \sqsubset t$) if t is a (resp. strict) extension of s. Suppose $A \subseteq \mathcal{S}_n$. If $s \in A$, we say that A *covers* s if for each t of length n which extends s, there exists $u \in A$ such that $s \sqsubset u \sqsubseteq t$. If A covers s, we write $c_A(s)$ for the set of $t \in A$ such that $s \sqsubset t$ and there does not exist any $u \in A$ such that $s \sqsubset u \sqsubset t$.

Finally, we say that a partial inductive method $f : A \to [0,1]$ is *dynamically coherent* if

$$\min_{t \in c_A(s)} f(t) \leq f(s) \leq \max_{t \in c_A(s)} f(t)$$

for all $s \in A$ such that A covers s. If $A = \mathcal{S}_n$, then A covers every $s \in \mathcal{S}_{n-1}$ and $c_A(s) = \{s0, s1\}$, so this is indeed a generalization of the previous definition of dynamic coherence. We are now able to state our main result.

Theorem 4.2. *Let $A \subseteq \mathcal{S}_n$ and $f : A \to [0,1]$. Then f extends to a speed-optimal convergent inductive method if and only if f is dynamically coherent.*

This result provides a completely general answer to the question whether arbitrary short-term behavior is compatible with speed-optimality: the partial inductive methods that have speed-

optimal extensions are exactly the dynamically coherent ones. In other words, speed-optimality induces dynamic coherence in the short-term. By strengthening Reichenbach's convergence criterion so that speed-optimality is required, one can avoid the objection that long-run requirements do not constrain short-term behavior. Dynamic coherence is necessary in the short-term if the long-run goal of speed-optimal convergence is to be achieved.

Before concluding this section, we address a question that arises naturally in view of the preceding results. Is there anything more precise to be said about the relation between rigidity and dynamic coherence? There is. Roughly, we will show that rigidity on a larger domain of binary sequences than $\mathcal{C}_{\text{conv}}$ is *equivalent* to dynamic coherence. In other words, the two concepts are equivalent given the right domain of definition.

To show this let us say that an inductive method ϕ is *dynamically coherent* if the restriction of ϕ to $\mathcal{S}_n$ is dynamically coherent for all n. That is, ϕ is dynamically coherent if for all $s \in \mathcal{S}$

$$\phi(s0) \leq \phi(s) \leq \phi(s1) \text{ or } \phi(s1) \leq \phi(s) \leq \phi(s0).$$

Since each instance of dynamic coherence involves only finitely many values of ϕ, Theorem 4.2 implies that if ϕ is a speed-optimal convergent inductive method, then ϕ is dynamically coherent.

Next, let $\mathcal{C}_\phi \subseteq \mathcal{C}$ be the set of binary sequences σ such that $(\phi(\sigma^n))_{n\in\mathbb{N}}$ converges to a limit. Note that $\mathcal{C}_\phi \supseteq \mathcal{C}_{\text{conv}}$ if ϕ is convergent. For all $s \in \mathcal{S}$, let

$$I^*_{\phi,s} = \bigcap_{\sigma\in\mathcal{C}_\phi} I_{\phi,s,\sigma},$$

and say that ϕ is *rigid** if $I^*_{\phi,s} = \{\phi(s)\}$ for all s. Note that if ϕ is convergent, then $I^*_{\phi,s} \subseteq I_{\phi,s}$ since $\mathcal{C}_\phi \supseteq \mathcal{C}_{\text{conv}}$.

Proposition 4.1. *A convergent inductive method is dynamically coherent if and only if it is rigid***.*

Since a convergent method is rigid* if it is rigid, Proposition 4.1 shows that we can view rigidity (or equivalently, speed-optimality) as nothing more than a minor strengthening of dynamic coherence. We do not currently know whether there are convergent inductive methods that are rigid* but not rigid. This leaves open the possibility that the two rigidity concepts are equivalent for convergent methods. In other words, it is an open question whether every convergent, dynamically

coherent inductive method is speed-optimal. We will discuss this open question more in the next section.

4.5 Discussion

In addition to discussing the problems that our analysis has left open, we would like to conclude by drawing some connections between our results and probabilistic learning. This will raise some interesting possibilities for future research.

One of the most distinguished proponents of dynamic coherence in settings where agents' degrees of beliefs are represented by probability measures is Brian Skyrms.[10] A key insight of Skyrms's work in this area is that dynamically coherent degrees of belief form *martingales*. Martingales, in turn, have especially nice convergence properties. Using this fact, Skyrms has shown that dynamic coherence implies (almost surely) convergent degrees of belief—and this holds quite generally, without the assumption that beliefs change by Bayesian conditionalization, for instance. An important philosophical consequence of this result, emphasized by Skyrms, is that dynamic coherence rules out a particularly strong kind of inductive skepticism (Skyrms, 2014).[11] If one's beliefs are dynamically coherent, and so convergent, then one must expect that one's own beliefs will exhibit regularities in the long-run. In the presence of coherence, absolute skepticism—the view that nature exhibits no regularities whatsoever—is untenable.

One question that Skyrms does not explicitly answer is whether dynamic coherence is necessary for convergent degrees of belief in the probabilistic setting. In fact, it is not. This follows from a small body of mathematical literature produced in the 1970s that, so far as we know, has never been mentioned in philosophical work on convergence and coherence.[12] This literature shows that degrees of belief are convergent in Skyrms's sense just in case they are *martingales in the limit*, a property strictly weaker than being a martingale. So convergent degrees of belief need not be dynamically coherent. In view of the fact that convergence for degrees of belief is a strictly weaker property than dynamic coherence, the following question arises naturally: Is there a compelling

[10] Skyrms (1987, 1990, 1996, 2006). Also see Huttegger (2013, 2014, 2015b, 2017b).

[11] Also see Diaconis and Skyrms (2017, ch. 10).

[12] In particular, this follows from results in Blake (1978). Also see Blake (1970); Mucci (1973, 1976); Edgar and Sucheston (1976, 1977).

notion of convergence for degrees of belief that strengthens Skyrms's notion and *does* imply dynamic coherence in the probabilistic sense? More specifically, is there a notion of speed-optimal convergence in the probabilistic setting that is sufficient for dynamic coherence? To the best of our knowledge, these questions are wide open.

These gaps in the probabilistic setting are, in a sense, dual to the ones that we have left open in our paper. To show this, we begin by remarking that there is a connection to be made with martingales in our framework as well. In our case, the relevant notion of martingale comes not from probability theory but the theory of algorithmic randomness. An inductive method ϕ is called a *martingale* if

$$\phi(s) = \frac{\phi(s0) + \phi(s1)}{2}$$

for all $s \in \mathcal{S}$. This notion of martingale was introduced by Jean Ville (1936; 1939) and plays an important role in contemporary studies of random binary sequences.[13]

It is clear from the definition that any inductive method that is a martingale is dynamically coherent in the sense of the previous section. As we also indicated in the previous section, our analysis has left open the question whether convergent martingales are necessarily speed-optimal. More generally, an important question for future research in our framework is: Are convergent, dynamically coherent inductive methods speed-optimal? Or, equivalently: Are convergent, rigid* methods rigid?

In the probabilistic setting, then, there is the question whether a notion of speed-optimal convergence is *sufficient* for coherence. And in the frequency prediction setting there is the question whether speed-optimal convergence is *necessary* for coherence. These questions, while independently interesting, are especially intriguing when considered together. We hope that future research will not only answer the open questions raised here but also shed light on unifying connections between induction in the probabilistic setting of Skyrms and induction in the frequency prediction setting of Reichenbach.

[13]Nies (2009). Also see Shafer and Vovk (2005) and Bienvenu et al. (2009).

Appendix

Proof of Theorem 4.1

First, suppose ϕ is a rigid convergent inductive method and ϕ' is any other convergent inductive method. Then $\phi(s) \neq \phi'(s)$ for some s. Since ϕ is rigid, there exists some $\sigma \in \mathcal{C}_{conv}$ extending s such that $\phi'(s) \notin I_{\phi,s,\sigma}$. It follows that $\epsilon_{\phi',s,\sigma} > \epsilon_{\phi,s,\sigma}$ and so ϕ' cannot be faster than ϕ.

Conversely, suppose ϕ is a convergent inductive method which is not rigid. Then for some s, $I_{\phi,s} = [a, b]$ is a nondegenerate interval.

Now let $t \in \mathcal{S}$ be a minimal finite extension of s such that $I_{\phi,s} \not\subseteq I_{\phi,t}$ (such a t exists since ϕ is convergent). Let u be t with its last bit removed; then u is also an extension of s.

Define

$$x = \begin{cases} \phi(t) & \phi(t) \in I_{\phi,s} \\ a & \phi(t) < a \\ b & \phi(t) > b. \end{cases}$$

Now define $\phi'(u) = x$ and $\phi'(v) = \phi(v)$ for all $v \neq u$. Note that since $x \in I_{\phi,s}$ and $I_{\phi,s} \subseteq I_{\phi,u}$ by minimality of t, $\phi' \not\prec_{\sigma,\epsilon} \phi$ for all $\sigma \in \mathcal{C}_{conv}$ and all $\epsilon > 0$. On the other hand, since $I_{\phi,s} \not\subseteq I_{\phi,t}$, there is some $\sigma \in \mathcal{C}_{conv}$ extending t such that $I_{\phi,s} \not\subseteq I_{\phi,t,\sigma}$. Note that $\phi(t) \in I_{\phi,t,\sigma}$. So, if $\phi(t) \in I_{\phi,s}$, then $\phi'(u) = \phi(t) \in I_{\phi,t,\sigma} = I_{\phi',t,\sigma}$ and so $I_{\phi',u,\sigma} = I_{\phi,t,\sigma} \not\supseteq I_{\phi,s}$. If $\phi(t) < a$, then $b \notin I_{\phi,t,\sigma}$ (otherwise $I_{\phi,t,\sigma}$ would contain all of $I_{\phi,s}$) and it follows that $b \notin I_{\phi',u,\sigma}$ since we defined $\phi'(u) = a$. Similarly, if $\phi(t) > b$, then $a \notin I_{\phi',u,\sigma}$.

So in all cases, we have $I_{\phi',u,\sigma} \not\supseteq I_{\phi,s}$, and in particular $I_{\phi',u,\sigma} \neq I_{\phi,u,\sigma}$ since $I_{\phi,u,\sigma} \supseteq I_{\phi,u} \supseteq I_{\phi,s}$. It follows that $I_{\phi',u,\sigma} \subset I_{\phi,u,\sigma}$, and therefore $\epsilon_{\phi',u,\sigma} < \epsilon_{\phi,u,\sigma}$. So, $\phi' \succ_{\sigma,\epsilon} \phi$ for $\epsilon = \epsilon_{\phi',u,\sigma}$. Thus ϕ' is faster than ϕ, and ϕ is not speed-optimal. □

Proof of Lemma 4.1

First, suppose that $\phi : \mathcal{S} \to [0, 1]$ is a convergent inductive method extending f and that f is not dynamically coherent. Let $s \in \mathcal{S}_{n-1}$ witness that f is incoherent. We assume that $f(s0) < f(s)$ and $f(s1) < f(s)$, as the other case is similar. Let $a = \max(f(s0), f(s1))$. Then $[a, f(s)] = [a, \phi(s)] \subseteq I_{\phi,s}$ since any $\sigma \in \mathcal{C}_{\text{conv}}$ extending s must have $\sigma^{|s|+1} \in \{s0, s1\}$. Hence, ϕ is not rigid, and by

Theorem 4.1, not speed-optimal.

Now suppose that f is dynamically coherent. Define $\phi : \mathcal{S} \to [0,1]$ by $\phi(s) = f(s)$ if $|s| \leq n$ and

$$\phi(s) = \frac{\sum_{i>n} s_i + f(s^n)}{|s| - n + 1}$$

if $|s| > n$. Then ϕ is convergent and extends f.

To prove that ϕ is rigid and hence speed-optimal, let $s \in \mathcal{S}$. If $|s| \geq n$, let $\sigma = s0000\ldots$ be the sequence obtained by extending s with all 0s. Then $\sigma \in \mathcal{C}_{conv}$ with $\ell(\sigma) = 0$ and the values $\phi(\sigma^m)$ are monotone decreasing for $m \geq |s|$. It follows that $I_{\phi,s,\sigma} = [0, \phi(s)]$. Similarly, if we take $\sigma' = s1111\ldots$, then $I_{\phi,s,\sigma'} = [\phi(s), 1]$. Thus $I_{\phi,s} \subseteq I_{\phi,s,\sigma} \cap I_{\phi,s,\sigma'} = \{\phi(s)\}$ and so $I_{\phi,s} = \{\phi(s)\}$.

In the case $|s| < n$, we use a similar argument but with different sequences. Since ϕ extends f and f is dynamically coherent, we can choose $a_1, a_2 \ldots, a_{n-|s|}$ such that

$$\phi(s) \geq \phi(sa_1) \geq \phi(sa_1a_2) \geq \cdots \geq \phi(sa_1a_2\ldots a_{n-|s|}).$$

Taking $\sigma = sa_1a_2\ldots a_{n-|s|}0000\ldots$, then we have as before that the values $\phi(\sigma^m)$ are monotone decreasing for $m \geq |s|$ and so $I_{\phi,s,\sigma} = [0, \phi(s)]$. Similarly, we can choose $b_1, b_2 \ldots, b_{n-|s|}$ such that

$$\phi(s) \leq \phi(sb_1) \leq \phi(sb_1b_2) \leq \cdots \leq \phi(sb_1b_2\ldots b_{n-|s|})$$

and then $\sigma' = sb_1b_2\ldots b_{n-|s|}1111\ldots$ satisfies $I_{\phi,s,\sigma'} = [\phi(s), 1]$. Thus again, we have $I_{\phi,s} = \{\phi(s)\}$. □

Proof of Theorem 4.2

The proof requires a preliminary lemma.

In the proof of the lemma, it will be convenient to say *f is coherent at s*, by which we mean that $f(s)$ does not witness a counterexample to dynamic coherence, as defined in the main text.

Lemma 4.2. *Let $A \subset \mathcal{S}_n$ and let $f : A \to [0,1]$ be dynamically coherent. Then there exists $s \in \mathcal{S}_n \setminus A$ and an extension $g : A \cup \{s\} \to [0,1]$ of f which is dynamically coherent.*

Proof. Let $s \in \mathcal{S}_n \setminus A$ be minimal with respect to extension. If s is not the empty string, let $t \in \mathcal{S}_n$

be such that (without loss of generality) $s = t0$. By minimality of s, we must have $t \in A$. Note that $A \cup \{s\}$ will not cover any elements of $\mathcal{S}_n$ not covered by A, except possibly t. Moreover, if A covers u and $u \neq t$, then $c_{A\cup\{s\}}(u) = c_A(u)$. So to check that an extension $g : A \cup \{s\} \to [0,1]$ of f is coherent, we need only check coherence at s and at t.

To define g and prove it is coherent, we consider several cases.

First, suppose s is not covered by A. In that case, we define $g(s) = f(t)$, or we define $g(s)$ arbitrarily if s is the empty string. In this case $A \cup \{s\}$ still will not cover s. If $A \cup \{s\}$ covers t, note that $s \in c_{A\cup\{s\}}(t)$ and so since $g(s) = g(t)$, s witnesses the coherence of g at t.

Now suppose s is covered by A. Let I be the closed interval spanned by $f(c_A(s))$. As long as we define $g(s)$ to be some element of I, then g will be coherent at s. So if s is the empty string, we just define $g(s)$ to be any element of I.

If s is not the empty string, first suppose that t is not covered by A. Since $s = t0$ is covered by A but t is not covered by A, there must be some u of length n extending $t1$ such that there is no $v \in A$ with $t < v \leq u$. But then this u witnesses that t is still not covered by $A \cup \{s\}$, so we may define $g(s)$ to be any element of I.

Finally, suppose t is covered by A. Since f is coherent at t, there exist $u, v \in c_A(t)$ such that $f(u) \leq f(t) \leq f(v)$. If u and v both extend $t1$, then $u, v \in c_{A\cup\{s\}}(t)$ as well, so we can define $g(s)$ to be any element of I. If u extends $t0$ and v extends $t1$, then $u \in c_A(s)$ and so $f(u) \in I$, so we can define $g(s) = f(u)$. Then g is coherent at t because $s, v \in c_{A\cup\{s\}}(t)$ and $g(s) \leq g(t) \leq g(v)$. Similarly if u extends $t1$ and v extends $t0$, we can define $g(s) = f(v)$. Finally, if u and v both extend $t0$, then $u, v \in c_A(s)$ and so $f(t) \in I$ since it is between $f(u)$ and $f(v)$. So, we may define $g(s) = f(t)$, and then g is coherent at t since $s \in c_{A\cup\{s\}}(t)$. □

Proof of Theorem 4.2. First, suppose $f : A \to [0,1]$ is dynamically coherent. If $A \neq \mathcal{S}_n$, then by Lemma 4.2, we can extend f to one more element of $\mathcal{S}_n$ while preserving its coherence. Iterating this, we may extend f to a partial inductive method $g : \mathcal{S}_n \to [0,1]$ which is coherent. By Lemma 4.1, we can then extend g to a speed-optimal convergent inductive method $\phi : \mathcal{S} \to [0,1]$.

Conversely, suppose f is not dynamically coherent, and let $s \in A$ witness the incoherence of f. Then A covers s and either $\min_{t\in c_A(s)} f(t) > f(s)$ or $\max_{t\in c_A(s)} f(t) < f(s)$. We write $a = \max_{t\in c_A(s)} f(t)$ and assume that $a < f(s)$ as the other case is similar.

Now suppose that ϕ is any convergent inductive method extending f. For any $\sigma \in \mathcal{C}_{conv}$ extending s, we have $\sigma^m \in c_A(s)$ for some $m > |s|$, since A covers s. We thus have $\phi(\sigma^m) = f(\sigma^m) \leq a$ for some $m > |s|$. It follows that $[a, f(s)] = [a, \phi(s)] \subseteq I_{\phi,s}$. Hence ϕ is not rigid, and by Theorem 4.1, not speed-optimal. □

Proof of Proposition 4.1

If a convergent inductive method $\phi : \mathcal{S} \to [0,1]$ is not dynamically coherent, then its restriction to some $\mathcal{S}_n$ is not dynamically coherent. The proof of Lemma 4.1, shows that ϕ is not rigid, but actually the same argument (applied to all $\sigma \in \mathcal{C}_\phi$ and not just all $\sigma \in \mathcal{C}_{conv}$) shows that ϕ is not rigid*.

Conversely, suppose a convergent inductive method $\phi : \mathcal{S} \to [0,1]$ is dynamically coherent. Let $s \in \mathcal{S}$; we wish to show $I^*_{\phi,s} = \{\phi(s)\}$. Since ϕ is dynamically coherent, there is some $a_1 \in \{0,1\}$ such that $\phi(sa_1) \leq \phi(s)$. There is similarly $a_2 \in \{0,1\}$ such that $\phi(sa_1a_2) \leq \phi(sa_1)$. Continuing by induction, we obtain a sequence σ extending s such that $\phi(\sigma^m) \leq \phi(\sigma^n)$ for all $m, n \geq |s|$ such that $m \geq n$.

Since the values $\phi(\sigma^m)$ form an eventually monotone sequence, they converge to some limit, so $\sigma \in \mathcal{C}_\phi$. Moreover, since $\phi(\sigma^m)$ is decreasing for $m \geq |s|$, the right endpoint of $I_{\phi,s,\sigma}$ is $\phi(s)$.

We may similarly construct a sequence σ' such that $\phi(\sigma'^m)$ is increasing for $m \geq |s|$ and so the left endpoint of $I_{\phi,s,\sigma'}$ is $\phi(s)$. Thus $I^*_{\phi,s} \subseteq I_{\phi,s,\sigma} \cap I_{\phi,s,\sigma'} = \{\phi(s)\}$ and so $I^*_{\phi,s} = \{\phi(s)\}$, as desired. □

Chapter 5

A Short Proof of the Blackwell-Dubins Theorem

The theorem of Blackwell and Dubins (1962) concerning the merging of conditional probabilities is a cornerstone in theorizing about rational learning. It has been widely discussed and applied by researchers in a number of fields including economics (Kalai and Lehrer, 1993, 1994; Pomatto et al., 2014), statistics (Diaconis and Freedman, 1986; D'Aristotile et al., 1988; Schervish and Seidenfeld, 1990), and philosophy (Earman, 1992; Huttegger, 2015b; Nielsen and Stewart, 2018; Stewart and Nielsen, 2019), and it has played a central role in this dissertation.

The purpose of this note is to provide a short and simple proof of a result that is slightly more general than the Blackwell-Dubins theorem. The added generality comes from relaxing assumptions that guarantee the existence of regular conditional probabilities (or "predictive" conditional probabilities in Blackwell and Dubins's terminology). Blackwell and Dubins made regularity an explicit assumption; Kalai and Lehrer (1994) work with a filtration that is generated by countable partitions, which in turn implies the existence of regular conditional probabilities. The result with regular conditional probabilities also appears as Proposition 10.4.25 in Bogachev (2007). In this note, we work with arbitrary conditional probabilities. Gaifman and Snir (1982) provided an independent proof of the merging of opinions theorem that uses the Cauchy-Schwarz inequality and a truncation argument. The approach here, on the other hand, contributes to the point of view that merging of opinions is really a part of martingale theory. I also provide a simple proof of the converse to the Blackwell-Dubins theorem, which was first observed by Kalai and Lehrer (1994, Theorem 2). The present proof of this result relies only on Fatou's lemma.

Let $(\Omega, \mathcal{F})$ be a measurable space equipped with a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ that increases to $\mathcal{F}$. Let P and Q be probability measures on $(\Omega, \mathcal{F})$, and, for all $A \in \mathcal{F}$ and $n \in \mathbb{N}$, let $P_n(A)$ (resp. $Q_n(A)$) be a version of $P(A \mid \mathcal{F}_n)$ (resp. $Q(A \mid \mathcal{F}_n)$), and let P^n (resp. Q^n) be the restriction of P (resp.

Q) to $\mathcal{F}_n$. Let $d_n = \operatorname{ess\,sup}_{A\in\mathcal{F}} |P_n(A) - Q_n(A)|$, where the essential supremum is with respect to Q.

Theorem 5.1. *Suppose that $Q^n \ll P^n$ for all $n \in \mathbb{N}$. Then, $d_n \to 0$ a.s. (Q) if and only if $Q \ll P$.*

Proof of Sufficiency. Assume $Q \ll P$ and let $q = dQ/dP$ be the corresponding Radon-Nikodym derivative. For all n, let q_n be a version of $\mathbb{E}_P(q \mid \mathcal{F}_n)$. By integrating both sides of the following equation over $F \in \mathcal{F}_n$ with respect to P we prove

$$Q_n(A)q_n = \mathbb{E}_P(\mathbf{1}_A q \mid \mathcal{F}_n) \text{ a.s. } (P/Q). \tag{5.1}$$

Note that $q_n > 0$ a.s. (Q). Using (5.1) we get

$$|P_n(A) - Q_n(A)| = \frac{|\mathbb{E}_P(\mathbf{1}_A(q_n - q) \mid \mathcal{F}_n)|}{q_n} \le \frac{\mathbb{E}_P(|q_n - q| \mid \mathcal{F}_n)}{q_n} \text{ a.s. } (Q). \tag{5.2}$$

As (5.2) holds for all $A \in \mathcal{F}$,

$$d_n \le \frac{\mathbb{E}_P(|q_n - q| \mid \mathcal{F}_n)}{q_n} \text{ a.s. } (Q). \tag{5.3}$$

Now, $q_n \to q$ a.s. (P/Q) and $q > 0$ a.s. (Q). So, by (5.3) it remains only to prove $\mathbb{E}_P(|q_n - q| \mid \mathcal{F}_n) \to 0$ a.s. (Q). Fix $\epsilon > 0$ and, for all n, let $B_n = \{q_n - q > \epsilon/2\}$. Note that, since $q_n \to q$ a.s. (P/Q),

$$P_n(B_n) \to 0 \text{ a.s. } (P/Q) \text{ and } Q_n(B_n) \to 0 \text{ a.s. } (Q), \tag{5.4}$$

for example by Theorem 5.5.9 in Durrett (2010). Then, by (5.4), and the vector lattice identity $x^+ = \frac{1}{2}(|x| + x)$,

$$\begin{aligned}
\mathbb{E}_P(|q_n - q| \mid \mathcal{F}_n) &= 2\mathbb{E}_P((q_n - q)^+ \mid \mathcal{F}_n) \\
&\le 2\mathbb{E}_P(\mathbf{1}_{B_n}(q_n - q) \mid \mathcal{F}_n) + \epsilon \\
&= 2q_n P_n(B_n) + 2q_n Q_n(B_n) + \epsilon \text{ a.s. } (Q).
\end{aligned}$$

Thus, by (5.4), $\limsup_n \mathbb{E}_P(|q_n - q| \mid \mathcal{F}_n) \le \epsilon$ a.s. (Q), and, since ϵ is arbitrary, the proof is complete. □

Proof of Necessity. Assume there exists $A \in \mathcal{F}$ such that $P(A) = 0$ while $Q(A) = \epsilon > 0$. Then, for all $n \in \mathbb{N}$, $P_n(A) = 0$ a.s. (P^n/Q^n), and

$$\mathbb{E}_Q(d_n) \geq \mathbb{E}_Q(Q_n(A)) = Q(A) = \epsilon \tag{5.5}$$

Let $M = \{d_n \to 0\}$. By (5.5),

$$\epsilon \leq \limsup_{n\to\infty} \mathbb{E}_Q(d_n \mathbf{1}_M) + \limsup_{n\to\infty} \mathbb{E}_Q(d_n \mathbf{1}_{M^c}). \tag{5.6}$$

By Fatou's lemma and (5.6) it follows that

$$\epsilon \leq \mathbb{E}_Q\Big[\limsup_{n\to\infty} d_n \mathbf{1}_M\Big] + \mathbb{E}_Q\Big[\limsup_{n\to\infty} d_n \mathbf{1}_{M^c}\Big] \leq Q(M^c)$$

Thus, $Q(M) \leq 1 - \epsilon$. □

Remark 5.1. If regular versions of the conditional probabilities for P are available, then, as in the proof of Blackwell and Dubins (1962), regular versions of the conditional probabilities for Q can be defined in a way that allows the essential supremum in the definition of d_n to be replaced by an ordinary supremum.[1]

[1]Many thanks to George Lowther for discussing this point with me and for helping me to correct an erroneous version of the proof of sufficiency. Any remaining errors are mine. George has also pointed out to me that Theorem 5.1 continues to hold, by the same argument, with $d_n = \operatorname{ess\,sup}_X |\mathbb{E}_P(X \mid \mathcal{F}_n) - \mathbb{E}_Q(X \mid \mathcal{F}_n)|$, where the essential supremum is over all random variables X taking values in $[0, 1]$ a.s. (P/Q).

Bibliography

Al-Najjar, N., L. Pomatto, and A. Sandroni (2014). Claim validation. *American Economic Review 104*(11), 3725–36.

Aliprantis, C. D. and K. C. Border (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer-Verlag, Berlin.

Autzen, B. (2018). Bayesian convergence and the fair-balance paradox. *Erkenntnis 83*(2), 253–263.

Belot, G. (2013). Bayesian orgulity. *Philosophy of Science 80*(4), 483–503.

Belot, G. (2017). Objectivity and bias. *Mind 126*(503), 655–695.

Bienvenu, L., G. Shafer, and A. Shen (2009). On the history of martingales in the study of randomness. *Electronic Journal for History of Probability and Statistics 5*(1), 1–40.

Billingsley, P. (2008). *Probability and Measure*. John Wiley & Sons.

Blackwell, D. and L. Dubins (1962). Merging of opinions with increasing information. *The Annals of Mathematical Statistics 33*(3), 882–886.

Blake, L. (1970). A generalization of martingales and two consequent convergence theorems. *Pacific Journal of Mathematics 35*(2), 279–283.

Blake, L. H. (1978). Every amart is a martingale in the limit. *Journal of the London Mathematical Society 2*(2), 381–384.

Bogachev, V. I. (2007). *Measure Theory*, Volume 2. Springer Science & Business Media.

Carnap, R. (1950). *Logical Foundations of Probability*. University of Chicago Press.

Carnap, R. (1952). *The Continuum of Inductive Methods*. University of Chicago Press.

Chen, R. (1977). On almost sure convergence in a finitely additive setting. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 37*(4), 341–356.

Christensen, D. (2009). Disagreement as evidence: The epistemology of controversy. *Philosophy Compass 4*(5), 756–767.

Cisewski, J., J. B. Kadane, M. J. Schervish, T. Seidenfeld, and R. Stern (2018). Standards for modest Bayesian credences. *Philosophy of Science 85*(1), 53–78.

D'Aristotile, A., P. Diaconis, and D. Freedman (1988). On merging of probabilities. *Sankhyā: The Indian Journal of Statistics, Series A 50*(3), 363–380.

Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association 77*(379), 605–610.

de Finetti, B. (1972). *Probability, induction, and statistics*. John Wiley & Sons.

de Finetti, B. (1974). *Theory of Probability*, Volume 1. John Wiley & Sons.

DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association 69*(345), 118–121.

Diaconis, P. and D. Freedman (1986). On the consistency of bayes estimates. *The Annals of Statistics 14*(1), 1–26.

Diaconis, P. and B. Skyrms (2017). *Ten Great Ideas about Chance*. Princeton University Press.

Doob, J. L. (1953). *Stochastic Processes*. New York: Wiley.

Dubins, L. E. (1975). Finitely additive conditional probabilities, conglomerability and disintegrations. *The Annals of Probability 3*(1), 89–99.

Dubins, L. E. and L. J. Savage (1965). *How to gamble if you must: Inequalities for stochastic processes*. Dover Publications.

Durrett, R. (2010). *Probability: Theory and Examples*. Cambridge University Press.

Earman, J. (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge, MA: MIT Press.

Easwaran, K. (2014). Regularity and hyperreal credences. *Philosophical Review 123*(1), 1–41.

Edgar, G. and L. Sucheston (1977). Martingales in the limit and amarts. *Proceedings of the American Mathematical Society 67*(2), 315–320.

Edgar, G. A. and L. Sucheston (1976). Amarts: A class of asymptotic martingales a. discrete parameter. *Journal of Multivariate Analysis 6*(2), 193–221.

Edwards, W., H. Lindman, and L. J. Savage (1963). Bayesian statistical inference for psychological research. *Psychological review 70*(3), 193–242.

Elga, A. (2016). Bayesian humility. *Philosophy of Science 83*, 305–323.

Elkin, L. and G. Wheeler (2018). Resolving peer disagreements through imprecise probabilities. *Noûs 52*(2), 260–278.

Epstein, L. G. and M. Le Breton (1993). Dynamically consistent beliefs must be Bayesian. *Journal of Economic theory 61*(1), 1–22.

Epstein, L. G. and M. Schneider (2003). Recursive multiple-priors. *Journal of Economic Theory 113*(1), 1–31.

Field, H. (1978). A note on Jeffrey conditionalization. *Philosophy of Science 45*(3), 361–367.

Gaifman, H. (2009). In A. Hájek and V. F. Hendricks (Eds.), *5 Questions: Probability and Statistics*, Chapter 4, pp. 41–57. Automatic Press/VIP.

Gaifman, H. (2013). The sure thing principle, dilations, and objective probabilities. *Journal of Applied Logic 11*(4), 373–385.

Gaifman, H. and M. Snir (1982). Probabilities over rich languages, testing and randomness. *The Journal of Symbolic Logic 47*(03), 495–548.

Gaifman, H. and A. Vasudevan (2012). Deceptive updating and minimal information methods. *Synthese 187*(1), 147–178.

Gallow, J. D. (2014). How to learn from theory-dependent evidence; or commutativity and holism: A solution for conditionalizers. *The British Journal for the Philosophy of Science 65*(3), 493–519.

Genest, C. and J. V. Zidek (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science 1*(1), 114–135.

Glymour, C. (1980). *Theory and Evidence*. Princeton University Press.

Goodman, N. (1955). *Fact, Fiction, and Forecast*. Harvard University Press.

Greaves, H. (2007). Probability in the everett interpretation. *Philosophy Compass 2*(1), 109–128.

Gyenis, Z. and M. Rédei (2017). General properties of Bayesian learning as statistical inference determined by conditional expectations. *The Review of Symbolic Logic 10*(4), 719–755.

Hájek, A. (2011). Staying regular. In *Australasian Association of Philosophy Conference*.

Halmos, P. R. (1950). *Measure Theory*. D. Van Nostrand Company, Inc.

Hegselmann, R., U. Krause, et al. (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of artificial societies and social simulation 5*(3).

Herron, T., T. Seidenfeld, and L. Wasserman (1994). The extent of dilation of sets of probabilities and the asymptotics of robust bayesian inference. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, Number 1, pp. 250–259. Philosophy of Science Association.

Herron, T., T. Seidenfeld, and L. Wasserman (1997). Divisive conditioning: Further results on dilation. *Philosophy of Science 64*(3), 411–444.

Howson, C. (2000). *Hume's problem: Induction and the justification of belief*. Clarendon Press.

Hume, D. (1748). *An Enquiry Concerning Human Understanding*. London: Millar.

Huttegger, S. M. (2013). In defense of reflection. *Philosophy of Science 80*(3), 413–433.

Huttegger, S. M. (2014). Learning experiences and the value of knowledge. *Philosophical Studies 171*(2), 279–288.

Huttegger, S. M. (2015a). Bayesian convergence to the truth and the metaphysics of possible worlds. *Philosophy of Science 82*(4), 587–601.

Huttegger, S. M. (2015b). Merging of opinions and probability kinematics. *The Review of Symbolic Logic 8*(04), 611–648.

Huttegger, S. M. (2017a). Analogical predictive probabilities. *Mind*.

Huttegger, S. M. (2017b). *The Probabilistic Foundations of Rational Learning*. Cambridge University Press.

Jeffrey, R. (1992). *Probability and the Art of Judgment*. Cambridge University Press.

Jeffrey, R. (2004). *Subjective Probability: The Real Thing*. Cambridge University Press.

Jern, A., K.-M. K. Chang, and C. Kemp (2014). Belief polarization is not always irrational. *Psychological review 121*(2), 206.

Johnson, W. E. (1924). *Logic, Part III: The Logical Foundations of Science*. Cambridge University Press.

Johnson, W. E. (1932). Probability: The deductive and inductive problems. *Mind 41*(164), 409–423.

Juhl, C. F. (1994). The speed-optimality of Reichenbach's straight rule of induction. *The British Journal for the Philosophy of Science 45*(3), 857–863.

Kadane, J. B. (2009). In A. Hájek and V. F. Hendricks (Eds.), *5 Questions: Probability and Statistics*, Chapter 9, pp. 97–114. Automatic Press/VIP.

Kadane, J. B., M. J. Schervish, and T. Seidenfeld (1996). Reasoning to a foregone conclusion. *Journal of the American Statistical Association 91*(435), 1228–1235.

Kalai, E. and E. Lehrer (1993). Rational learning leads to Nash equilibrium. *Econometrica 61*(5), 1019–1045.

Kalai, E. and E. Lehrer (1994). Weak and strong merging of opinions. *Journal of Mathematical Economics 23*(1), 73–86.

Kelly, K. T. (1996). *The Logic of Reliable Inquiry*. Oxford University Press.

Kelly, K. T., O. Schulte, and C. Juhl (1997). Learning theory and the philosophy of science. *Philosophy of Science 64*(2), 245–267.

Kelly, T. (2008). Disagreement, dogmatism, and belief polarization. *The Journal of Philosophy 105*(10), 611–633.

Kolmogorov, A. N. (1950). *Foundations of the Theory of Probability*. Chelsea Publishing Company.

Lane, D. A. and W. D. Sudderth (1984). Coherent predictive inference. *Sankhyā: The Indian Journal of Statistics, Series A: The Indian Journal of Statistics, Series A 46*(2), 166–185.

Lane, D. A. and W. D. Sudderth (1985). Coherent predictions are strategic. *The Annals of Statistics 13*(3), 1244–1248.

Lehrer, K. and C. Wagner (1981). *Rational Consensus in Science and Society: A Philosophical and Mathematical Study*, Volume 21. Dordrecht, Holland: D. Reidel Publishing Company.

Leitgeb, H. (2017). Imaging all the people. *Episteme 14*(4), 463–479.

Levi, I. (1980). *The Enterprise of Knowledge*. MIT Press, Cambridge, MA.

Levi, I. (1982). Conflict and social agency. *The Journal of Philosophy 79*(5), 231–247.

Levi, I. (1985a). Consensus as shared agreement and outcome of inquiry. *Synthese 62*(1), pp. 3–11.

Levi, I. (1985b). Imprecision and indeterminacy in probability judgment. *Philosophy of Science 52*(3), 390–409.

Levi, I. (2009). Why indeterminate probability is rational. *Journal of Applied Logic 7*(4), 364–376.

Lévy, P. (1937). *Théorie de l'addition des variables aléatoires*. Gauthiers-Villars, Paris.

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *The Philosophical Review 85*, 3.

Lewis, D. (1980). A subjectivist's guide to objective chance. In W. L. Harper, R. Stalnaker, and G. Pearce (Eds.), *Ifs: Conditionals, Belief, Decision, Chance, and Time*, pp. 267–297. Dordrecht, Holland: D. Reidel Publishing Company.

Lindley, D. V. (2006). *Understanding uncertainty*. John Wiley & Sons.

Lipecki, Z. (2001). Cardinality of the set of extreme extensions of a quasi-measure. *manuscripta mathematica 104*(3), 333–341.

Lord, C. G., L. Ross, and M. R. Lepper (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology 37*(11), 2098.

Miller, R. I. and C. W. Sanchirico (1999). The role of absolute continuity in "merging of opinions" and "rational learning". *Games and Economic Behavior 29*(1-2), 170–190.

Mucci, A. G. (1973). Limits for martingale-like sequences. *Pacific Journal of Mathematics 48*(1), 197–202.

Mucci, A. G. (1976). Another martingale convergence theorem. *Pacific Journal of Mathematics 64*(2), 539–541.

Nielsen, M. (2018). Deterministic convergence and strong regularity. *The British Journal for the Philosophy of Science*.

Nielsen, M. and R. T. Stewart (2018). Persistent disagreement and polarization in a Bayesian setting. *The British Journal for the Philosophy of Science*.

Nies, A. (2009). *Computability and randomness*. Oxford University Press.

Pedersen, A. P. and G. Wheeler (2014). Demystifying dilation. *Erkenntnis 79*(6), 1305–1342.

Pedersen, A. P. and G. Wheeler (2015). Dilation, disintegrations, and delayed decisions. In *Proceedings of the 9th International Symposium on Imprecise Probability: Theories and Applications*, pp. 227–236.

Peirce, C. S. (1992a). The fixation of belief. In N. Houser and C. Kloesel (Eds.), *The Essential Peirce, Volume 1: Selected Philosophical Writings (1867–1893)*, pp. 109–123. Indiana University Press.

Peirce, C. S. (1992b). How to make our ideas clear. In N. Houser and C. Kloesel (Eds.), *The Essential Peirce, Volume 1: Selected Philosophical Writings (1867–1893)*, pp. 124–141. Indiana University Press.

Plachky, D. (1976). Extremal and monogenic additive set functions. *Proceedings of the American Mathematical Society 54*(1), 193–196.

Pomatto, L., N. Al-Najjar, A. Sandroni, et al. (2014). Merging and testing opinions. *The Annals of Statistics 42*(3), 1003–1028.

Pomatto, L. and A. Sandroni (2018). An axiomatic theory of inductive inference. *Philosophy of Science 85*(2), 293–315.

Popper, K. R. (1955). Two autonomous axiom systems for the calculus of probabilities. *The British Journal for the Philosophy of Science 6*(21), 51–57.

Purves, R. A. and W. D. Sudderth (1976). Some finitely additive probability. *The Annals of Probability 4*(2), 259–276.

Purves, R. A. and W. D. Sudderth (1983). Finitely additive zero-one laws. *Sankhyā: The Indian Journal of Statistics, Series A 45*(1), 32–37.

Ramsey, F. P. (1931). Truth and probability. In R. B. Braithwaite (Ed.), *The Foundations of Mathematics and Other Essays*, pp. 156–198. Kegan, Paul, Trench, Trubner, & Co.

Reichenbach, H. (1938). *Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge*. University of Chicago press.

Reichenbach, H. (1949). *The Theory of Probability*. University of California Press.

Renyi, A. (1970). *Foundations of Probability*. Holden Day.

Rosenthal, J. S. (2006). *A First Look at Rigorous Probability Theory*. World Scientific.

Ross, L. and C. A. Anderson (1982). Shortcomings in the attribution process: On the origins and maintenance of erroneous social assessments. In D. Kahneman and A. Tversky (Eds.), *Judgment Under Uncertainty: Heuristics and Biases*, Chapter 9, pp. 129–152. New York: Oxford Univeristy Press.

Royden, H. L. and P. Fitzpatrick (2010). *Real analysis* (4th ed.). Pearson Education.

Salmon, W. C. (1966). *The Foundations of Scientific Inference*. University of Pittsburgh Press.

Salmon, W. C. (1991). Hans reichenbach's vindication of induction. *Erkenntnis 35*(1-3), 99–122.

Savage, L. (1954). *The Foundations of Statistics*. Wiley, New York.

Savage, L. (1972). *The Foundations of Statistics*. New York: John Wiley & Sons.

Schechter, E. (1996). *Handbook of Analysis and its Foundations*. Academic Press.

Schervish, M. and T. Seidenfeld (1990). An approach to consensus and certainty with increasing evidence. *Journal of Statistical Planning and Inference 25*(3), 401–414.

Schervish, M. J., T. Seidenfeld, and J. B. Kadane (1984). The extent of non-conglomerability of finitely additive probabilities. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 66*(2), 205–226.

Schirokauer, O. and J. B. Kadane (2007). Uniform distributions on the natural numbers. *Journal of Theoretical Probability 20*(3), 429–441.

Schulte, O. (1999a). The logic of reliable and efficient inquiry. *Journal of Philosophical Logic 28*(4), 399–438.

Schulte, O. (1999b). Means-ends epistemology. *The British Journal for the Philosophy of Science 50*(1), 1–31.

Seidenfeld, T. (2001). Remarks on the theory of conditional probability: Some issues of finite versus countable additivity. In V. F. Hendricks, S. A. Pedersen, and K. F. Jørgensen (Eds.), *Probability Theory*, pp. 167–178. Dordrecht, The Netherlands: Klewer.

Seidenfeld, T., J. B. Kadane, and M. J. Schervish (1989). On the shared preferences of two Bayesian decision makers. *The Journal of Philosophy 86*(5), 225–244.

Seidenfeld, T., M. J. Schervish, and J. B. Kadane (2010). Coherent choice functions under uncertainty. *Synthese 172*(1), 157–176.

Seidenfeld, T. and L. Wasserman (1993). Dilation for sets of probabilities. *The Annals of Statistics 21*(3), 1139–1154.

Shafer, G. and V. Vovk (2005). *Probability and finance: it's only a game!*, Volume 491. John Wiley & Sons.

Shimony, A. (1955). Coherence and the axioms of confirmation. *The Journal of Symbolic Logic 20*(01), 1–28.

Shogenji, T. (1999). Is coherence truth conducive? *Analysis 59*(264), 338–345.

Skyrms, B. (1980). *Causal Necessity*. New Haven: Yale Academic Press.

Skyrms, B. (1987). Dynamic coherence and probability kinematics. *Philosophy of Science 54*(1), 1–20.

Skyrms, B. (1990). *The dynamics of rational deliberation*. Harvard University Press.

Skyrms, B. (1995). Strict coherence, sigma coherence and the metaphysics of quantity. *Philosophical Studies 77*(1), 39–55.

Skyrms, B. (1996). The structure of radical probabilism. *Erkenntnis 45*(2-3), 285–297.

Skyrms, B. (2006). Diachronic coherence and radical probabilism. *Philosophy of Science 73*(5), 959–968.

Skyrms, B. (2014). Grades of inductive skepticism. *Philosophy of Science 81*(3), 303–312.

Stewart, R. T. and M. Nielsen (2019). Another approach to consensus and maximally informed opinions with increasing evidence. *Philosophy of Science 86*(2), 1–19.

Stewart, R. T. and I. Ojea Quintana (2018). Probabilistic opinion pooling with imprecise probabilities. *Journal of Philosophical Logic 47*(1), 17–45.

Sunstein, C. R. (2002). The law of group polarization. *Journal of political philosophy 10*(2), 175–195.

Suppes, P. (1966). A Bayesian approach to the paradoxes of confirmation. *Studies in Logic and the Foundations of Mathematics 43*, 198–207.

van Fraassen, B. C. (1984). Belief and the will. *The Journal of Philosophy 81*(5), 235–256.

van Fraassen, B. C. (1999). Conditionalization, a new argument for. *Topoi 18*(2), 93–96.

van Fraassen, B. C. (2000). The false hopes of traditional epistemology. *Philosophical and Phenomenological Research LX*(2), 253–280.

van Fraassen, B. C. and J. Y. Halpern (2016). Updating probability: Tracking statistics as criterion. *The British Journal for the Philosophy of Science 68*(3), 725–743.

Ville, J. (1936). Sur la notion de collectif. *Comptes rendus des Séances de l'Académie des Sciences 203*, 26–27.

Ville, J. (1939). *Etude critique de la notion de collectif*. Gauthier-Villars Paris.

Wagner, C. (2002). Probability kinematics and commutativity. *Philosophy of Science 69*(2), 266–278.

Wagner, C. (2003). Commuting probability revisions: The uniformity rule. *Erkenntnis 59*(3), 349–364.

Wasserman, L. and T. Seidenfeld (1994). The dilation phenomenon in robust Bayesian inference. *Journal of Statistical Planning and Inference 40*(2), 345–356.

Weatherson, B. (2015). For Bayesians, rational modesty requires imprecision. *Ergo 2*.

White, R. (2005). Epistemic permissiveness. *Philosophical perspectives 19*(1), 445–459.

White, R. (2010). Evidential symmetry and mushy credence. *Oxford studies in epistemology 3*(161), 20.

Williamson, T. (2002). *Knowledge and its Limits*. Oxford University Press.

Zabell, S. L. (2002). It all adds up: The dynamic coherence of radical probabilism. *Philosophy of Science 69*(S3), S98–S103.

Zimper, A. and A. Ludwig (2009). On attitude polarization under bayesian learning with non-additive beliefs. *Journal of risk and uncertainty 39*(2), 181–212.