DETERMINISTIC CONVERGENCE AND STRONG REGULARITY

SUBMISSION FOR THE DAVID H. SIFF PHILOSOPHY OF SCIENCE AWARD

Abstract. Bayesians since Savage (1954) have often appealed to asymptotic results to counter charges of excessive subjectivity. Their claim is that objectionable differences in prior probability judgments will vanish as agents learn from evidence, and individual agents will converge to the truth. 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 under which the updated probabilities of two agents actually converge to the same settled opinion. I call this mode of convergence *deterministic convergence*. The main results extend those found in Huttegger (2015b). The present results 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.

1. Introduction and Overview

Classical Bayesian epistemology is often criticized for being too subjective. Bayesian theory requires only that inquirers have internally coherent "prior" probabilities (de Finetti, 1990) and that these probabilities be updated by conditionalizing on evidence. Because it is rationally permissible, on the Bayesian 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 certain mild assumptions, the probabilities of two Bayesian agents who learn the same sequence of propositions are *almost sure* to eventually reach a consensus (Savage, 1954, 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,

Date: July 10, 2017.

[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, 1954, p. 67).

1992; Huttegger, 2015a,b; Weatherson, 2015; Elga, 2016). What they show, essentially, is that as evidence is gathered differences between prior probability assignments are "washed out" and intersubjective agreement is achieved, with probability 1. The theorems purport to show that the widespread disagreement that Bayesianism permits, which some find objectionable, is merely a transient phenomenon.[3]

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 John Earman (1992, Chapter 6) and has been developed more recently by Gordon Belot (2013; 2016).[4]

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 for readers who are not interested in the (inevitably) technical details.

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 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*). 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,

[3]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).

[4]Haim Gaifman has voiced similar concerns. About the convergence 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).

Bayesians may wish to relinquish notions of objectivity altogether and adopt a thoroughgoing subjectivism.

The rest of the paper proceeds as follows. The next section introduces the mathematical tools that we will need to study deterministic convergence. In Section 3, we state and discuss the main results of the paper. Section 4 addresses an objection to the approach taken in Section 3. In Section 5, we provide an additional characterization result for Bayesian updates and present the argument for strong regularity. Section 6 contains concluding remarks.

2. Mathematical Preliminaries

We need only three mathematical objects: probability spaces, filtrations on probability spaces, and random variables. I will begin by discussing the mathematical structure of these objects and how they are used in Bayesian theory. Then I will discuss the convergence and merging theorems.

A *probability space* is a triple $(\Omega, \mathcal{F}, P)$ consisting of an arbitrary set Ω, a sigma-algebra $\mathcal{F}$ of subsets of Ω and a probability measure P defined on $\mathcal{F}$. The members of Ω are called *points* or *possible worlds*. A sigma-algebra $\mathcal{F}$ is a non-empty collection of subsets of Ω that is closed under countable set operations (unions, intersections, complements, symmetric differences, and so on). The members of $\mathcal{F}$ are called (measurable) *events* or *propositions*. A probability measure P is a real-valued, nonnegative, countably additive set function on $\mathcal{F}$ that assigns the sure event Ω measure 1, i.e. $P(\Omega) = 1$. To say that P is countably additive means that if F is a countable union of pairwise disjoint events $\{F_i : i \in \mathbb{N}\}$, then the probability of F is equal to the sum of the probabilities of the F_i. If P assigns probability 1 to an event F, i.e. $P(F) = 1$, then we say that F occurs *almost surely*.

A Bayesian agent is represented by a probability space $(\Omega, \mathcal{F}, P)$, and P is called her *prior*. The prior is updated by *conditionalization* on learned evidence. This model of learning consists of three assertions. (1) Each rational learning experience corresponds to a learned proposition E, thought of as the agent's newly acquired evidence. (2) Rational learning requires raising one's probability in the proposition E to 1. (3) 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, these assertions can be expressed succinctly by writing

$$P_E(A) = P(A \mid E) := \frac{P(A \cap E)}{P(E)},$$

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, which is a ratio, is not defined when $P(E) = 0$. This means that Bayesian 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 state the asymptotic results that were mentioned in the previous section, we require a modest generalization of conditionalization. It is at this point that our second object, a filtration, appears. Let $(\Omega, \mathcal{F})$ consist of, as above, a set Ω and a sigma-algebra of subsets of Ω. A *sigma-subalgebra* of $\mathcal{F}$ is a subset of $\mathcal{F}$ that is itself a sigma-algebra of subsets of Ω. A *filtration* $\{\mathcal{F}_n : n \in \mathbb{N}\}$ on $(\Omega, \mathcal{F})$ is a sequence of increasing sigma-subalgebras of $\mathcal{F}$, that is each $\mathcal{F}_n$ is a subset of $\mathcal{F}_{n+1}$. Symbolically, $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$ for all $n \in \mathbb{N}$.

A filtration on $(\Omega, \mathcal{F})$ represents the information that an agent learns at time n. Think of Ω as a set of possible worlds, one of which is actual. The agent is uncertain which ω in

Ω is the actual world and her uncertainty is represented by her prior probability measure P over $(\Omega, \mathcal{F})$. Learning a member $\mathcal{F}_n$ of a filtration simply involves assigning each proposition in $\mathcal{F}_n$ a probability, and these probability assignments may differ from the agent's prior assignments. At each time n, the information in $\mathcal{F}_n$ fixes a posterior probability assignment over the members of $\mathcal{F}_n$. That is all that learning involves for us.

This model of learning generalizes standard conditionalization, according to which the agent learns a single proposition E rather than an entire sigma-subalgebra of propositions. In the conditionalization model, each evidence proposition E generates a binary partition of Ω that separates the worlds at which E is true from the worlds at which E is false. And each partition generates a sigma-subalgebra of Ω by collecting together all possible countable unions of the partition's cells.[5] So we can recover standard conditionalization from the more general filtration model by considering sigma-subalgebras that are generated by events. The filtration model is also more general in the sense that we no longer require that learning involve assigning a proposition probability 1.

For our purposes, we can think of each sigma-subalgebra $\mathcal{F}_n$ as a (finite) partition, corresponding, perhaps, to the possible outcomes of an experiment that the agent performs at time n. Because filtrations are increasing, the agent acquires more information through time, ever refining the partition that represents the distinctions she is able to make amongst possible worlds. Mathematically, we needn't assume that $\mathcal{F}_n$ is generated by a partition, but thinking this way can serve as a useful heuristic. Each filtration generates a sigma-algebra, which we can think of as the total information contained in the filtration. In what follows we will assume that $\{\mathcal{F}_n : n \in \mathbb{N}\}$ generates $\mathcal{F}$. This means that, in the limit, the agent learns everything about the space of possibilities $(\Omega, \mathcal{F})$ on which her prior is defined.

Finally, a *random variable* is a real-valued function on Ω.[6] We think of random variables simply as numbers that depend on which world in Ω is actual. For example, let the random variable X represent the lowest daily average temperature in New York, in March 2018, measured in Fahrenheit. In some possible world ω_1, this temperature is 15, in some other world ω_2 it is 20, and so on. We write $X(\omega_1) = 15$ and $X(\omega_2) = 20$.

The concept of a random variable is essential to understanding the Bayesian convergence and merging theorems, to which we now turn. These theorems model an agent's posterior probabilities as random variables, the values of which are uncertain with respect to the agent's prior. The random variables that represent the agent's updated probabilities are generalizations of the conditional probabilities that were discussed above. It is not important, for our purposes, to dwell on the mathematical details of these more general conditional probabilities, but I include some discussion in a footnote.[7] The only point I wish to stress is that the values that these objects take depend on $\omega \in \Omega$, unlike the more elementary conditional probabilities mentioned above. Let us denote by $P^n(A)$ an agent's posterior probability of event A after learning $\mathcal{F}_n$. This is a function of Ω, so we may have, for example, in some possible world ω_1, $P^n(A)(\omega_1) = 0.5$, and in another possible world ω_2, $P^n(A)(\omega_2) = 0.2$.

[5]The sigma-subalgebra generated by E is simply $\{\varnothing, E, E^c, \Omega\}$, where E^c is the complement of E.

[6]Random variables X must be measurable as well: for all Borel subsets B of $\mathbb{R}$ the set $\{\omega : X(\omega) \in B\}$ must be a member of $\mathcal{F}$.

[7]The conditional probabilities $P(\cdot \mid \mathcal{F}_n)$ that are needed are conditional probabilities *given a sigma-subalgebra* $\mathcal{F}_n$. The existence of these objects is guaranteed by the Radon-Nikodym theorem, though they are not in general guaranteed to be well-behaved probabilities. For a summary of some of the issues involved see Seidenfeld (2001) or Huttegger (2015b). For a mathematical introduction to conditional probabilities given a sigma-subalgebra, I recommend Billingsley (2008).

The first convergence result that we will discuss is the so-called *convergence to the truth theorem*. The truth about a proposition A depends on which world is the actual one. We can represent this using the *indicator random variable* $\mathbf{1}_A$ *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 random variable returns 1 if the proposition A is true at the world ω and 0 if A is false at ω. We want to consider the event that the sequence $\{P^n(A) : n \in \mathbb{N}\}$ of posterior probabilities of A converges to $\mathbf{1}_A$ as n gets very large. This event can be represent as

$$\{\omega \in \Omega : P^n(A)(\omega) \to \mathbf{1}_A(\omega) \text{ as } n \to \infty\}.$$

To avoid messy notation, let us call this event C_A, the event that the agent converges to the truth about A. Now, the convergence to the truth theorem states that

$$P(C_A) = 1.$$[8]

That is, for each proposition A, a Bayesian agent with probabilities given by P assigns probability 1 to the event that her updated probabilities converge to the truth about A. Put another way, for each A, Bayesian agents almost surely converge to the truth about A. Although in general there are possible worlds at which convergence fails to occur, Bayesian agents always assign these worlds probability 0.

Even in possible worlds where convergence to the truth fails, however, a result of Blackwell and Dubins (1962) guarantees that, so long as two agents' priors assign probability 0 to the same events, and provided they update on the same increasing stream of evidence, they will almost surely achieve intersubjective agreement in the limit.[9] This *merging of opinions theorem* is the second result that we discuss. We now consider two Bayesian agents with priors P and Q defined on the same 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 distance that is independent of a particular event A, we define the 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 respect to events:

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

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

Let us suppose that P and Q are learn the same filtration and that the updated probabilities are denoted, as above, by P^n and Q^n respectively. Let us also suppose that P and Q do not differ too dramatically in their prior 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 posterior probability of A above 0. If the Q posterior probabilities of A tend to some positive value as n gets large, then P and Q cannot achieve agreement in the long run. Under the assumption of mutual

[8] This result 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.

[9] This result was discovered independently, in a somewhat different framework and with a different proof, by Gaifman and Snir (1982).

absolute continuity, the merging of opinions theorem states that almost surely

$$d(P^n, Q^n) \to 0 \text{ as } n \to \infty.$$ [10]

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 accumulates.[11] In the limit, P and Q will achieve consensus (almost surely).

I close this section with some brief remarks about the absolute continuity condition in the merging of opinions theorem. 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 coincides with 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.

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).[12]

The fact is that Bayesian learning does not guarantee convergence to the truth, or asymptotic 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. In view of Earman's objection, it is natural to ask under what conditions the troublesome almost surely qualifications can be removed from the Bayesian convergence results. It is our aim in the remainder of this paper to answer this question.

To that end, we must depart somewhat from the model of Bayesian learning that was explained in the previous section. In particular, we will no longer regard updated probabilities as random variables about which an agent has certain probabilistic *beliefs*. Rather, we want to know under what conditions *actual* Bayesian 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 space $(\Omega, \mathcal{F})$ having some additional

[10]The converse to this theorem also holds and was shown by Kalai and Lehrer (1994). Namely, if the distance between P^n and Q^n tends to 0 almost surely (for P and Q), and P and Q are mutually absolutley continuous for events in the filtration $\{\mathcal{F}_n : n \in \mathbb{N}\}$, then P and Q are mutually absolutely continuous for all events. In my presentation of the merging of opinions theorem I have, for economy of exposition, omitted discussion of the (necessary) requirement that P^n and Q^n be *regular conditional distributions*, or, in Blackwell and Dubins's terminology *predictive* conditional probabilities. See the references in footnote 2.

[11]Schervish and Seidenfeld (1990) extend the Blackwell and Dubins theorem, showing results in which sets of conditionalized probabilities merge uniformly.

[12]Similar points have been raised by Glymour (1980); Kelly et al. (1997); Howson (2000).

structure. The relevant mathematical structure is called a *topology*. Using topological methods, Gordon Belot (2013, 2016), 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).[13] 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. 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. (2017). 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. 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, we mean that there exists a *decreasing sequence* $\{E_n : n \in \mathbb{N}\}$ of events with positive prior probability, i.e. $E_1 \supseteq E_2 \supseteq ...$, and $P_n = P(\cdot \mid E_n)$ for all n.[14] Recall that this is a special case of learning a filtration, as discussed in Section 2. The definition of a sequence of updates does not depend on the presence of a filtration, but filtrations are important for the results below.

We now make a few more heuristic remarks, sharpening those made in Section 2, about the model of learning being developed. At each time n, an agent learns a member $\mathcal{F}_n$ of the filtration $\{\mathcal{F}_n : n \in \mathbb{N}\}$ and fixes her posterior probabilities P_n on members of $\mathcal{F}_n$. This could occur, for example, as the result of observing an experimental outcome. Next, the agent extends P_n to all of $\mathcal{F}$ via some sort of learning rule. In the Bayesian case, the agent learns some proposition E_n, sets $P_n(E_n) = 1$ and extends to all of $\mathcal{F}$ via the equation $P_n(A) = P(A \mid E_n)$. In what follows, except when we restrict ourselves to the Bayesian case, we needn't assume any particular extension procedure. The model is completely general in this respect.

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* to the set function P_∞ on $(\Omega, \mathcal{F})$ if

$$P_\infty(A) = \lim_{n\to\infty} P_n(A) \text{ for all } A \in \mathcal{F}.$$

[13] The argument also appears in Belot (2016).

[14] The assumption that the sequence $\{E_n : n \in \mathbb{N}\}$ is decreasing can be made without loss of generality. If this assumption fails for E_n and E_{n+1}, we simply replace E_{n+1} with $E_{n+1} \cap E_n$.

The function P_∞ is called the *deterministic limit* of the sequence $\{P_n : n \in \mathbb{N}\}$.[15] 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. 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 settle down to *some* stable limit. Brian Skyrms (1996) has called this convergence to a "maximally informed opinion."

A few remarks are in order about the generality of the framework that we are working in, 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[16]), 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 model of learning. In order to introduce the first condition, we now equip $(\Omega, \mathcal{F})$ with a filtration $\{\mathcal{F}_n : n \in \mathbb{N}\}$ that increases to $\mathcal{F}$. The first condition states that later learning experiences never contradict earlier ones. Suppose an agent is to learn a sequence of refining partitions. Then, we require that, once probabilities have been determined over a partition at time n, these values remain unchanged at time $n + 1$. Mathematically, we can write, for all n,

$$P_{n+1}(F) = P_n(F) \text{ for all } F \in \mathcal{F}_n.\text{[17]} \tag{M}$$

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.[18]

(M) requires that learned information is never lost or destroyed. In learning scenarios in which (M) fails, an agents probabilities for some event $F \in \mathcal{F}_n$ are free to oscillate indefinitely,

---

[15] In mathematics, this mode of convergence is called *setwise*.

[16] These alternative rules include Jeffrey conditioning (or probability kinematics) (Jeffrey, 1992) and various parameterizations thereof (Field, 1978; Wagner, 2002), imaging (Lewis, 1976; Leitgeb, 2016), and Gallow's rule for learning theory-dependent evidence (Gallow, 2014).

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

[18] Some have found this feature of Bayesian learning objectionable and have developed alternative accounts (Levi, 1980; Jeffrey, 1992; Williamson, 2002).

and she need not converge. Brian Skyrms (1996) has given a justification of (M) in terms of diachronic coherence. And Simon Huttegger (2014; 2015b) has suggested 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 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 given by $P(\omega_n) = 2^{-n}$, $n \in \mathbb{N}$. 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 $P_n(A) \geqslant \epsilon$.

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 $P_n(E_n) = 1 \geqslant 1/2$. 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 4) 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 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.[19] 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. *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 6). 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

[19] We remark that Huttegger shows deterministic convergence, but does not argue that the limiting set function is a probability measure. In fact, it is, as we point 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.

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). The proof can be found in the Appendix.

Theorem 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 $P_\infty \ll P$.*

Combining the last two results we have a complete characterization of deterministic convergence in terms of uniform absolute continuity.

Theorem 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 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. *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 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 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, $A \cap E_n = \varnothing$. 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.[20] In this example, the probabilities $P_n(A_0)$ oscillate forever and never reach a stable limit. △

[20]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 wants 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 5, we will see that further problems arise for Bayesians in the former group as well.

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 differences between different agents' probabilities 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 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.

Fortunately, 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, provided their priors do not disagree too drastically. 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 issues in the next section.

We will assume that two agents learn the same evidence if they update on the same filtration, and, at each time n, their probabilities over $\mathcal{F}_n$ are the same. Let P and Q be the probabilities of two distinct agents, defined on the same probability space. Formally, we require that, for all n, and all $A \in \mathcal{F}_n$,

$$P_n(A) = Q_n(A),$$

and we say that $\{P_n : n \in \mathbb{N}\}$ and $\{Q_n : n \in \mathbb{N}\}$ *learn the same evidence*. This is a natural way to formalize the notion of learning the same evidence in the current framework. At each time n, both agents receive the same information and they agree about that information's impact on their posterior probability assignments. 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).

In order to assume that P and Q learn the same evidence, we must also assume that P and Q are mutually absolutely continuous. 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 do not learn the same evidence. It should come as no surprise that mutual absolute continuity plays an important role here given its centrality in the original merging of opinions theorem of Blackwell and Dubins (1962).

The next result states that, under the assumptions just made and those of the previous section, the updates of any two mutually absolutely continuous probabilities that learn the same evidence converge to the same limiting distribution.[21]

Theorem 4. *Let P and Q be mutually absolutely continuous probability measures on the same probability space $(\Omega, \mathcal{F})$. 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 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 $\{Q_n : n \in \mathbb{N}\}$.*

The proof is given in the Appendix. Even in our very general framework, we see that disagreement vanishes as evidence accumulates. Given that the framework of deterministic convergence is able to capture a notion of consensus in the limit, I submit that the framework is readily applicable to the issues that were used to motivate the paper.

5. Strong Regularity

In Examples 1 and 2 of Section 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. It turns out that 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. Again, the proof of this result is given in the Appendix.

Theorem 5. *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) that, for updates by conditionalization, uniform absolute continuity is equivalent to deterministic convergence. Combining this result with Theorem 5 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 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.*

[21] Note that the the conclusion of the present result is somewhat weaker than a deterministic analogue of the Blackwell and Dubins merging of opinions theorem (Section 2). We do not show that $d(P_n, Q_n) \to 0$, but merely that the deterministic limits of the sequence of updates of P and the sequence of updates of Q are the same. Indeed, it seems that establishing a stronger merging-type result is impossible without further assumptions, though establishing this mathematically is outside the scope of the present paper. For a deterministic merging result for learning by probability kinematics, see Huttegger (2015b, Theorem 5.1)

(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 $\varnothing$. Equivalently, a probability is regular if it assigns strictly 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 (Lewis, 1980; Skyrms, 1980; Hájek, 2011; Easwaran, 2014). One of the main arguments 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 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 acquiesce in 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 strictly positive probability is countable, and hence P must assign zero probability to uncountably many $\{\omega\}$.[22] If probabilities are to satisfy 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 $\{x\}$ is Lebesgue measurable and has measure 0. Moreover, by countable additivity, every countable

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

subset of the unit interval is Lebesgue measurable with measure 0. There are also uncountable subsets of the unit interval of Lebesgue measure zero: for example, the Cantor set. Similar observations hold for any probability distribution that is absolutely continuous with respect to Lebesgue measure. Such distributions include, for example, the normal (or Gaussian) distribution, the log-normal distribution, the Beta and Gamma distributions, and many other commonly used probability distributions. 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 for the possibility that probabilities take their values in a hyperreal field containing 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 different 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.[23]

Finally, we briefly mention another fairly mainstream response to Lewis's argument for regularity, which is to reject the view that rational learning always goes by conditionalization. Several alternatives to conditionalization are mentioned in footnote 16.

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 sequence of decreasing 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.[24] But, as we have already seen, there are probabilities that are regular and not strongly regular. In Example 1 (Section 3), 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. 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 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) When rational agents learn a filtration and (M) holds, their updated probabilities converge deterministically.
(C2) Therefore (by Corollary 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, the filtration condition and (M) are satisfied. By (P6), the agent's updated probabilities converge deterministically. But this contradicts the lack of strong regularity, by Corollary 2. Therefore, rational probabilities are strongly regular.

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

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

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 *non-finite countable* spaces in the following fashion. 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 $P(\{\omega\}) \geqslant 1/n$. If this were not the case, then there would exist (as in Example 1) a decreasing sequence of events whose probabilities were not bounded away from zero.[25] 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.[26] Given an increasing stream of evidence (a filtration) that eventually reveals everything about the space of possibilities 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 settle down in the long run. The hope is simply that, in the highly idealized learning scenarios that we have been considering, 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. It is 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 5 that conditional probabilities be defined as ratios of unconditional probabilities. I do not currently know 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 has been endorsed by some Bayesians. 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

[25]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 $n \to \infty$.

[26]For a recent endorsement of something like (P6), see Autzen (2017).

accounts of objectivity and intersubjective agreement, then either Bayesians must settle for the almost sure qualifications that Earman, Belot, and others have attacked, or they must embrace the consequences of imposing deterministic convergence, including strong regularity.

6. Conclusion

To summarize, this paper's contributions have been, first, to raise a natural question about Bayesian asymptotics that has received relatively little attention in the philosophical literature, namely: Under what conditions to probabilistic updates converge deterministically? Second, we have provided a simple, but conclusive, answer to that question at a fairly 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—if only in highly idealized learning scenarios—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 view that rational degrees of belief are probabilistically coherent, and the diachronic component is the view 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 6 (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 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 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.

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 nonnegative 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 $F \in \mathcal{F}_n$,

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

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 nonnegative P-integrable function, P_∞ defined by $P_\infty(A) = \int_A Z_\infty dP$ is a measure on $\mathcal{F}$. Hence, $\{P_n : n \in \mathbb{N}\}$ converges deterministically to P_∞ on $\bigcup_n \mathcal{F}_n$. It remains to show that $\{P_n : n \in \mathbb{N}\}$ converges deterministically to P_∞ on all of $\mathcal{F}$. This last part of the proof is straightforward but somewhat technical.

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

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

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 $A \in \mathcal{F}$,

$$P(A) < \delta \ \text{ implies } \ P_n(A) < \epsilon/2.$$

For sufficiently large K we have

$$P\big(A - \bigcup_{k=1}^{k} A_k\big) \leqslant \delta \ \text{ and } \ P_\infty\big(A - \bigcup_{k=1}^{K} A_k\big) \leqslant \epsilon/2.$$

Then,

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

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,

$$\limsup_{n\to\infty} |P_n(A) - P_\infty(A)| \leqslant \epsilon.$$

This shows that $A \in \mathcal{C}$ because ϵ is arbitrary. We can now conclude that $\{P_n : n \in \mathbb{N}\}$ converges deterministically to P_∞ on all of $\mathcal{F}$.

We finish by observing that P_∞ is clearly a probability measure that is absolutely continuous with respect to P. □

Proof of Theorem 4

Proof. Given the assumptions, we know from Theorem 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 $n \geqslant n_0$. Hence, $P_n(A) = Q_n(A)$ if $n \geqslant n_0$, 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 5.

Proof. First, suppose it is not the case that $\alpha := \inf\{P(E_n) : n \in \mathbb{N}\} > 0$. Then, since in general $\alpha \geqslant 0$, 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 $P_n(A) \geqslant \epsilon$. 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, $P_n(E_n) = 1 \geqslant 1/2$ 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 $\delta \leqslant \epsilon P(E_n)$ 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

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

as desired. □

References

Autzen, B. (2017). Bayesian convergence and the fair-balance paradox. *Erkenntnis*, 1–11.

Belot, G. (2013). Bayesian orgulity. *Philosophy of Science 80*(4), 483–503.

Belot, G. (2016). Objectivity and bias. *Mind*.

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.

Cisewski, J., J. B. Kadane, M. J. Schervish, T. Seidenfeld, and R. Stern (2017). Standards for modest Bayesian credences. *Philosophy of Science*, Forthcoming.

de Finetti, B. (1990). *Theory of Probability*, Volume 1. John Wiley and Sons.

Dubins, L. E. (1975). Finitely additive conditional probabilities, conglomerability and disintegrations. *The Annals of Probability*, 89–99.

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.

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.

Field, H. (1978). A note on jeffrey conditionalization. *Philosophy of Science*, 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. and M. Snir (1982). Probabilities over rich languages, testing and randomness. *The journal of symbolic logic 47*(03), 495–548.

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.

Glymour, C. (1980). *Theory and Evidence.* Princeton University Press.

Hájek, A. (2011). Staying regular. In *Australasian Association of Philosophy Conference.*

Howson, C. (2000). *Hume's problem: Induction and the justification of belief.* Clarendon Press.

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.

Jeffrey, R. (1992). *Probability and the Art of Judgment.* Cambridge University Press.

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.

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.

Leitgeb, H. (2016). Imaging all the people. *Episteme* (DOI: 10.1017/epi.2016.14).

Levi, I. (1980). *The Enterprise of Knowledge.* MIT Press, Cambridge, MA.

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*, pp. 267–297. Springer.

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.

Renyi, A. (1970). *Foundations of probability.*

Royden, H. L. and P. Fitzpatrick (2010). *Real analysis* (4th ed.). Pearson Education.

Savage, L. (1972, originally published in 1954). *The Foundations of Statistics.* New York: John Wiley and Sons.

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.

Seidenfeld, T. (2001). Remarks on the theory of conditional probability: Some issues of finite versus countable additivity. In V. F. Hendricks (Ed.), *Probability Theory*, pp. 167–178. Klewer.

Skyrms, B. (1980). *Causal Necessity.* New Haven: Yale Academic Press.

Skyrms, B. (1996). The structure of radical probabilism. *Erkenntnis 45*(2-3), 285–297.

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.

Weatherson, B. (2015). For Bayesians, rational modesty requires imprecision. *Ergo, an Open Access Journal of Philosophy 2.*

Williamson, T. (2002). *Knowledge and its Limits.* Oxford University Press.