Journal of Philosophical Logic
https://doi.org/10.1007/s10992-025-09814-6

Updating by Maximizing Expected Accuracy in Infinite Non-Partitional Settings

Kenny Easwaran[1] · **Michael Nielsen**[2]

Received: 13 January 2025 / Accepted: 15 September 2025

Abstract

Greaves, H., & Wallace (*Mind*, 115(459), 607–632 2006) justify Bayesian conditionalization as the update plan that maximizes expected accuracy, for an agent considering finitely many possibilities, who is about to undergo a learning event where the potential propositions that she might learn form a partition. In recent years, several philosophers have generalized this argument to less idealized circumstances. Some authors (Easwaran, *Thought: A Journal of Philosophy*, 2(1), 53–61 2013; Nielsen, *Statistics & Probability Letters*, 185, 109412 2022) relax finiteness, while others (Carr, *Synthese*, 198(4), 3581–3601 2021; Gallow, *Noûs*, 55(3), 487–516 2021; Isaacs & Russell, 2022; Schultheis, 2023) relax partitionality. In this paper, we show how to do both at once. We give novel philosophical justifications of the use of σ-algebras in the infinite setting, and argue for a different interpretation of the "signals" in the non-partitional setting. We show that the resulting update plan mitigates some problems that arise when only relaxing finiteness, but not partitionality, such as the Borel-Kolmogorov paradox.

Keywords Expected accuracy · Conditionalization · Externalism · Non-partitional evidence · Infinite sigma-algebras · Formal epistemology · Borel-Kolmogorov paradox

1 Introduction

Starting at least with Joyce [22], a particular topic of research in formal epistemology has concerned the way in which various epistemic procedures can be justified entirely on the basis of accuracy. This paper does not seek to justify or critique this project, but just shows how it can be applied to one specific question.

In particular, Greaves and Wallace [16] show that, if an agent distributes their credence over only finitely many possible worlds, and the agent expects to receive a signal drawn from some partition, then standard Bayesian conditionalization is the

✉ Kenny Easwaran
easwaran@gmail.com

1 University of California, Irvine, CA 92697, USA

2 Texas A&M University, College Station, TX 77840, USA

Published online: 01 October 2025

update plan that maximizes expected accuracy. The point of this paper is to show how to relax those two substantive assumptions — we generalize to infinite sets of worlds, and signals that don't form a partition. We explain the details of their finite partitional setting, including both the substantive assumptions of theirs that we keep, and those we eventually relax, in Section 2.

We have each done partial generalizations of this work to the infinite setting [10, 29]. We summarize this work in Section 3. Importantly, in this paper we give additional philosophical justification for some of the formal tools that those papers just presupposed, particularly the use of countably-generated σ-algebras to represent important aspects of content.

Starting at least with Williamson [39] epistemologists have worried that partitionality builds in a substantive assumption that the agent not only comes to know something, but also comes to know *that* she knows it. Several authors have tried to generalize the expected accuracy framework to accommodate these non-partitional cases. In Section 4, we particularly discuss the proposals of [6, 14, 19, 33], and [34]. We provide a unified framework encompassing their proposals, though there may be some controversy over whether we have done so adequately. In particular, we describe updates as being driven by a "signal", which plays the roles of propositions, guesses about propositions, guesses about worlds, and other entities, for these authors. We think that most of their proposals can be seen as special cases of ours, though we leave out the agent's credences about her own credences, which are an important focus for some of these authors. (Readers who are just interested in this discussion can skip the mathematical technicalities of Section 3.) The moves we make here raise some worries about how to interpret the significance of maximizing expected accuracy, but we leave some of this interpretive work for future research.

Finally, in Section 5, we show how to put the work of Sections 3 and 4 together to determine what update plan maximizes expected accuracy in the infinite non-partitional setting. Interestingly, the non-partitionality allows for a partial resolution of the Borel-Kolmogorov paradox, which arises in the infinite, partitional setting. We describe several examples to show how our framework works. In particular, there are some examples that we suggest are best understood by allowing "signals" to be some type of entity that are distinct from propositions or worlds, and instead convey some type of "experiential" knowledge that the agent is not in a position to express in terms of worlds and truth-conditions.

2 The Greaves and Wallace Argument

Greaves & Wallace [16] assume (as do we, and all the other authors whose work we generalize) the basic framework of probability theory. That is, there is a set Ω of "worlds", and a set $\mathcal{F}$ of subsets of Ω whose members can be thought of as the propositions that the agent grasps, and a probability function $p\colon \mathcal{F} \to \mathbb{R}$ that represents the agent's credences. For p to be a probability function is for it to be non-negative on all its inputs, to assign the output 1 to the input Ω, and for it to satisfy countable additivity. That is, if there is a finite or countably infinite collection of disjoint elements of $\mathcal{F}$ whose union is also a member of $\mathcal{F}$, then the values assigned to each of them by p

must add up to the value assigned to their union. (Debates about the status of countable additivity are outside the scope of this paper.)

Greaves and Wallace assume that Ω is finite, and thus $\mathcal{F}$ is as well. We eventually relax this assumption in Section 3. But in this finite setting, it is also reasonable to assume that $\mathcal{F}$ is the entire power set of Ω, as the rest of this paragraph will argue. First, it is natural to assume that $\mathcal{F}$ is an algebra — that is, the complement of any element of $\mathcal{F}$ is also in $\mathcal{F}$, and so is the intersection or union of any two elements of $\mathcal{F}$. Furthermore, although we refer to the members of Ω as "worlds", it is clear that these usually aren't possible worlds in the sense that many philosophers think of them. In any normal sense of "possible world", there are infinitely many different possible worlds that make infinitely fine distinctions of various sorts, and ruling out all but finitely many of them would require unusually precise evidence. So the finite set of "worlds" in Ω must represent the limited conceptual distinctions the agent actually draws. If there is some $\omega \in \Omega$ whose singleton is not in the agent's algebra $\mathcal{F}$, then we can replace ω by the set $\bar{\omega}$ that is intersection of all elements of $\mathcal{F}$ containing ω. (Since $\mathcal{F}$ is finite, this intersection is finite, and is thus a member of $\mathcal{F}$.) In this way, it is natural to assume that every singleton subset of Ω is in $\mathcal{F}$. But since we have assumed that finite unions of members of $\mathcal{F}$ are in $\mathcal{F}$, and there are only finitely many members of Ω, this means that every subset of Ω is in $\mathcal{F}$.

Since we have argued that, in this context, $\mathcal{F}$ is the set of all subsets of Ω, one might prefer to use power set notation from set theory to represent it. However, in Section 3.1, we will argue that $\mathcal{F}$ should not generally be the power set of Ω in the infinite setting. Thus, we use the notation "$\mathcal{F}$" even in the finite setting, to keep the notation constant.

2.1 Signals and Plans

The setting of the Greaves and Wallace argument involves an agent who has some prior probability p and is about to learn something. Greaves and Wallace assume that learning occurs by an "experiment", in which the agent receives one of "a set of mutually exclusive and jointly exhaustive alternatives". They refer to an update plan as an "act" that yields a particular probability function in the case of any result they could receive from the experiment. Their goal is to identify the update plan with highest expected accuracy.

To represent this, we let $\Delta(\mathcal{F})$ be the set of all probability functions defined on the members of $\mathcal{F}$. We let S be the set of all "signals" the agent might receive. In the setting of Greaves and Wallace, these are the possible outcomes of the experiment, though we generalize this in Section 4. We assume that an update plan is a function $P : S \to \Delta(\mathcal{F})$. That is, for any $s \in S$, $P(s)$ will be a probability function, representing the credences in various propositions that the plan recommends in response to signal s.

It is unrealistic to think that an agent explicitly considers these plans, either to choose among them, or to evaluate them, or even to adopt one in particular. However, it is natural to suppose that the agent does have *some* disposition to update their credences in light of the signal they receive in the learning situation, and the point of this framework is to identify which disposition would be best, from the point of view of accuracy. If the signal represents all the information that the agent has to work with, (we will say more about what is involved in this assumption in Sections 3.1 and 4.3) then this disposition must assign a probability function to each such signal, and thus must be a function from S to $\Delta(\mathcal{F})$. The point of the project we are engaged in is to identify which such function is the best from the perspective of expected accuracy, whether or not agents can or do consciously evaluate them.

Note that we use capital "P" for an update plan, and lower case "p" for an individual probability function. A probability function is defined on propositions, so if $A \in \mathcal{F}$, then $p(A)$ will be some real number. Since $P(s)$ is itself a probability function, we should formally write $P(s)(A)$ for the probability value the plan recommends for proposition A in light of an update on signal s. But to avoid these double parentheses, we will generally use "P_s" to refer to the function $P(s)$, so that we can write "$P_s(A)$" instead of "$P(s)(A)$". It will also occasionally be helpful to refer to the function $P^A \colon S \to \mathbb{R}$ defined by $P^A(s) = P(s)(A)$, so that "P^A" refers to the part of the agent's plan that determines how her credence in A (separate from any other proposition) depends on the signal she receives.

We represent Greaves and Wallace's assumption that S is "a set of mutually exclusive and jointly exhaustive alternatives" by assuming, in this section, that each member of S is a proposition (i.e., a member of $\mathcal{F}$), and that each member of Ω is contained in exactly one member of S. A set of propositions such that every member of Ω is contained in exactly one of them is called a *partition*, and thus we refer to Greaves and Wallace as working in the "partitional" setting. In Section 4, we will discuss generalizations to learning situations that are not partitional. But in the partitional setting, the idea is that the signal the agent receives is the unique true member of S. We can represent this by letting L be the "learning" function $L \colon \Omega \to S$ that assigns to each $\omega \in \Omega$ the unique member of S that is true in ω (i.e., that contains ω).

Greaves and Wallace themselves express things slightly differently. Instead of thinking of an update plan as a function from *signals* to probability functions, they think of it as a function from *worlds* to probability functions, but require that it be "available", meaning that it assigns the same value (i.e., the same probability function) to any worlds that are members of the same element of the partition. As we will see in Section 4.3, this stipulation will cause problems in non-partitional settings. But we think that it is conceptually better to understand the update plan as being driven by the signal, rather than being driven by the world in a way that is constant with respect to something about the signal. While the "world" represents *all* the distinctions that the agent conceptualizes, the signal represents just the distinctions that the agent gets in the learning situation. In both the infinite and the non-partitional setting, the conceptualization of the "signal" may be very different from that of the world. However, we still can represent the role of the world in the partitional setting, by composing P with L. $P(L(\omega))$ is the result of updating according to plan P with learning function L in world ω.

2.2 Expected Value

The way we evaluate plans will involve their expected accuracy. Before we define this, it is necessary to define expected value more generally.

Let $f : \Omega \to \mathbb{R}$ be any function assigning a real number to each world. (In the terminology of mathematical probability theory, f is a *random variable*, but this terminology is often more confusing than enlightening.) Let p be any probability function on $\mathcal{F}$. In the finite setting, we define the expected value of f with respect to p, $Exp_p(f)$, as follows:

$$Exp_p(f) = \sum_{\omega \in \Omega} p(\{\omega\}) f(\omega).$$

That is, the expected value of a function is the sum of the value of that function at each world, multiplied by the probability of that world. This is naturally thought of as something like the estimate that someone with probability function p would have of the value of f — at any rate, it is the weighted average of the possible values of f, weighted by the probabilities.

Because Ω is finite, and the probability function is additive, this formula can be re-written in another way. Since there are only finitely many possible outputs of the function, we can group together like terms, and see that

$$Exp_p(f) = \sum_{x \in \mathbb{R}} x \cdot p(\{\omega \in \Omega : f(\omega) = x\}).$$

That is, rather than starting with each world, and considering the value the function takes there, and multiplying this by the probability of the world, we can start with each value the function takes on, and consider the probability of the set of all worlds at which it takes the value, and multiply those together. This second formulation is sometimes easier to work with (since there are often fewer terms in the sum), and plays an important role in Section 3.3, where we define expectations when Ω is infinite.

2.3 Accuracy

We need to define the concept of “accuracy” if we want to update by maximizing expected accuracy. We won’t say too much about what precisely we mean by this, because this concept is shared across the literature we are responding to (starting at least with [22], though there are older discussions in the statistical literature discussing a related concept under the term “scoring rule”, including [27]). The basic idea is that “accuracy” is a measure of how good or bad it is to have a given credence function in a given world, where this goodness or badness is a measure of how close the credences are to the truth. We represent accuracy as a function

$$U : \Delta(\mathcal{F}) \times \Omega \to \mathbb{R}.$$

This function takes as inputs a probability function and a world, and $U(p, \omega)$ represents the accuracy of having credence function p in world ω. Although “accuracy” and “scoring rule” are the most common names for this function, we already use the letters “A” and “S” for other purposes, so we represent this function with the letter “U”, for “epistemic utility”.

We will abuse notation slightly and let $U(p)\colon \Omega \to \mathbb{R}$ be the function that associates each world with the accuracy of having probability function p in that world — that is, $U(p)(\omega) = U(p, \omega)$. Because $U(p)$ is a function assigning real numbers to members of Ω, we can calculate the expected value of it. The one substantive constraint we assume on the accuracy function is:

Strict Propriety: For probability functions $p \neq q$, $Exp_p(U(p)) > Exp_p(U(q))$.

The basic idea of this principle is that any probability function should expect itself to be more accurate than it expects any other probability function to be. That is, an agent who has some probability function shouldn't judge some other probability function to be, unconditionally, more accurate than their own. There have been various criticisms of Strict Propriety in the literature (for instance, [4, 31]), but it is assumed in the set of results that we are responding to, so we will not question it here.

Since this is the only substantive constraint we apply to the accuracy function, it may be helpful to point out that this constraint does in fact partially cash out what it means to be an *accuracy* function. In particular, for any world ω there is an "omniscient" probability function p_ω given by $p_\omega(A) = 1$ iff $\omega \in A$ and $p_\omega(A) = 0$ otherwise. We can see that $Exp_{p_\omega}(U(p)) = U(p, \omega)$, so the instance of Strict Propriety with $p = p_\omega$ tells us that $U(p_\omega, \omega) > U(q, \omega)$, when q is any other probability function. The omniscient function at a given world must be the highest-scored function in that world. Similarly, out of the probability functions that give all probability to two worlds, we can show that a Strictly Proper accuracy function gives one of these functions higher accuracy in one of these worlds iff it gives a higher probability to that world. None of this is sufficient to show that Strict Propriety cashes out the full *meaning* of the concept of accuracy, but it is at least suggestive.

As some examples of accuracy functions, we can consider two standard ones. The quadratic accuracy function is given by

$$U_{quad}(p, \omega) = -\sum_{A \in \mathcal{F}} (p(A) - p_\omega(A))^2.$$

To show that it satisfies Strict Propriety, one can calculate that $Exp_p(U_{quad}(p)) - Exp_p(U_{quad}(q)) = \sum_{A \in \mathcal{F}} (p(A) - q(A))^2$, which is strictly positive unless $p = q$. The logarithmic accuracy function is given by

$$U_{log}(p, \omega) = \log p(\{\omega\}).$$

To show that it satisfies Strict Propriety, one calculates that

$$Exp_p(U_{log}(q)) = \sum_{\omega \in \Omega} p(\{\omega\}) \log q(\{\omega\}).$$

The derivative of this expectation with respect to a small change in $q(\{\omega\})$ is $p(\{\omega\})/q(\{\omega\})$. If there are some ω with $q(\{\omega\}) > p(\{\omega\})$ and some with $q(\{\omega\}) < p(\{\omega\})$ (which always occurs, unless $p = q$), then taking away a small amount of probability from the former to give to the latter will result in a net increase in expected accuracy.

2.4 The Result

Now that we have the concepts of an update plan, accuracy, and expected value, we can show how these are put together to yield the expected accuracy of an update plan. Recall that $L\colon \Omega \to S$ is the function showing which signal (from the partition) is learned at a world, and $P\colon S \to \Delta(\mathcal{F})$ is the update plan determining which probability function to update to in light of a given signal.

If we want to estimate the expected accuracy of updating according to plan P, we can note that $P(L(\omega))$ is the probability function the agent would end up with if they were actually in world ω, so that they would receive signal $L(\omega)$. So the accuracy they would receive from updating with plan P in world ω is $U(P(L(\omega)), \omega)$. We can abuse notation and let "$U(P)$" stand for the function that takes ω to $U(P(L(\omega)), \omega)$.

We then want to identify which plan P maximizes $Exp_p(U(P))$, when p is the agent's prior probability distribution. In order to do this, we first define an auxiliary probability function p_s by $p_s(A) = p(A \cap s)/p(s)$, whenever $s \in S$ and $p(s) > 0$. This is a probability function — and in particular, it is the one recommended by conditionalization on s.

Now we can chase some definitions to get the equation needed for the central result. Note that $Exp_p(U(P)) = \sum_{\omega\in\Omega} p(\{\omega\})U(P)(\omega)$. We can break this up as $\sum_{s\in S}\sum_{\omega\in s} p(\{\omega\})U(P)(\omega)$, because each ω is in exactly one s. Then we can observe that $p(s)p_s(\{\omega\}) = p(\{\omega\})$ if $\omega \in s$, and 0 otherwise. Thus, the inner sum can be extended to all of $\omega \in \Omega$ if we replace $p(\{\omega\})$ by $p(s)p_s(\{\omega\})$. Thus, $Exp_p(U(P)) = \sum_{s\in S}\sum_{\omega\in\Omega} p(s)p_s(\{\omega\})U(P)(\omega)$. Since $p(s)$ doesn't depend on ω, we can factor it out of the sum over ω, yielding $Exp_p(U(P)) = \sum_{s\in S} p(s)\sum_{\omega\in\Omega} p_s(\{\omega\})U(P)(\omega)$. Note also that for $\omega \in s$ (the only ones for which $p_s(\{\omega\})$ is non-zero), $U(P)(\omega) = U(P_s, \omega)$, because for any $\omega \in s$, P updates to P_s. Thus, we get:

$$Exp_p(U(P)) = \sum_{s\in S} p(s)Exp_{p_s}(U(P_s)).$$

In this last equation, we see the important thing. Each update P_s in light of a different signal s contributes a separate term to the overall expected accuracy of P. We can maximize the expected accuracy of P by maximizing the contribution of each term P_s. Because each term contributes through a non-negative multiple of $Exp_{p_s}(U(P_s))$, and because of Strict Propriety, the expected accuracy is maximized iff for each s, $P_s = p_s$; that is, the plan for updating in light of s, equals the result of conditionalizing p on s. Thus, in the finite partitional setting, conditionalization is the unique plan that maximizes expected accuracy.

Now that we have shown how the argument works in the finite partitional setting, we can start to generalize it. In Section 3, we remove the assumption that Ω is finite, while in Section 4, we remove the assumption that S forms a partition of Ω. In Section 5, we then show how the argument works when both assumptions are removed. However, we still keep the basic structure, that Ω is the set of "worlds" for the agent, that $\mathcal{F}$ is the set of propositions over it, that p is a countably-additive probability function on $\mathcal{F}$, that S is the set of potential signals the agent can respond to, that a plan P sends each signal to a probability function, that there is a measure U of accuracy, and that we want to find a plan P that maximizes expected accuracy. In the infinite case, we make changes prompted by thinking more carefully about $\mathcal{F}$, while in the non-partitional case we think more carefully about S.[1]

3 Letting Ω be Infinite

When the set Ω of possibilities is infinite, it is standard not to assume that every subset forms a proposition that the agent grasps, and that not every subset has a probability. Instead, it is standard to assume that $\mathcal{F}$, the set of propositions that have a probability, forms a σ-algebra. That means, for any countable collection of members of $\mathcal{F}$, their union and their intersection are also in $\mathcal{F}$, as is the complement of any individual member of $\mathcal{F}$. To assume that the collection of propositions that an agent directly grasps forms a σ-algebra would be a significant idealization. However, this can be somewhat weakened. We will describe how in Section 3.1. Then, in Section 3.2, we will extend the idea of grasping a proposition to grasping a function. This gives rise to the notion of a measurable function. In Section 3.3, we briefly explain how to define expected values in infinite settings. In Section 3.4, we explain what it means for accuracy to be measurable, and we generalize the definition of Strict Propriety in a straightforward way. In Section 3.5, we explain how to define a σ-algebra on the set of signals. Once we have all of these concepts, we can finally, in Section 3.6, set up the problem of updating by maximizing expected accuracy in infinite settings. In that section we prove a generalization of the Greaves-Wallace result, and we show how it makes sense of the famous Borel-Kolmogorov paradox.

3.1 Grasping a σ-algebra

If X is some collection of subsets of Ω, we say the σ-algebra *generated* by X, $\sigma(X)$, is the smallest σ-algebra that contains every member of X.[2] We say that a set F is a *base* for a σ-algebra $\mathcal{F}$ iff F is a subset of $\mathcal{F}$, and there are no two distinct probability functions on $\mathcal{F}$ that agree on their values for all elements of F. We don't need to think of the agent directly grasping every member of $\mathcal{F}$ in order to define probabilities on all

[1] Here and throughout the paper, we follow the formalism that has grown in the philosophical literature around Greaves and Wallace's approach. A referee for this journal has pointed out another formalism built around the work of [2, 3] that approaches related problems. We have not done the translation between the two approaches but believe it is largely compatible.

[2] $\sigma(X)$ can be defined "top-down" by defining it as the intersection of all σ-algebras that contain X. $\sigma(X)$ can also be defined "bottom-up" by letting X_0 be the set of all members of X and complements of members of X, and for any countable ordinal α, letting X_α be the collection of all countable unions and countable intersections of members of X_β with $\beta < \alpha$. $\sigma(X)$ is then the union of all X_α. This bottom-up definition shows that if X is countable, then each X_i has at most the cardinality of the continuum, $2^{\aleph_0}$, so $\sigma(X)$ has cardinality at most $\aleph_1 \cdot 2^{\aleph_0}$, which is the same as the continuum.

members of $\mathcal{F}$ — it suffices if the agent directly grasps each member of some base for $\mathcal{F}$ and has a probability for each of them, and then the rest are in a sense "implicitly grasped", since their probabilities are uniquely determined by the probabilities of the propositions the agent directly grasps.

Importantly, the Carathéodory extension theorem states that if F is an algebra of sets (not necessarily a σ-algebra — an algebra is just closed under *finite* unions and intersections, as well as complements), then F forms a base for $\sigma(F)$. While it may be implausible to think that the collection of propositions directly grasped by an agent forms a σ-algebra, it is plausible to think that it forms an algebra (as we assumed already in Section 2), and we can think of $\mathcal{F}$ as the σ-algebra generated by the propositions the agent directly grasps. Then, the probabilities of the propositions the agent directly grasps uniquely determine the probabilities of every proposition in the full σ-algebra generated by them. It is thus reasonable to say that the agent "implicitly" grasps the members of $\mathcal{F}$, even if they only directly grasp the members of some *base* for it.[3]

To illustrate these points, consider a case where someone is about to look at a dial, or an unmarked clock, where the pointer could be at any of a continuum of positions around the edge. We can identify Ω with the set of positions that the pointer could be at, and name these positions as ω_φ, where $0 \leq \varphi < 2\pi$. It is natural to assume that each singleton $\{\omega\}$ is in $\mathcal{F}$.[4] By $\mathcal{F}$'s properties as a σ-algebra, every countable subset of Ω is also in $\mathcal{F}$, as is the complement of any countable set (such a set is called "co-countable"). Let $\mathcal{C}$ be this σ-algebra consisting of all and only the countable and co-countable subsets of Ω.

It's very plausible to think that an ordinary agent would be able to grasp more propositions than just those in $\mathcal{C}$. In particular, it is natural to assume that every interval $\{\omega_\varphi : \varphi_0 < \varphi < \varphi_1\}$ with $0 \leq \varphi_0 < \varphi_1 < 2\pi$ is also in $\mathcal{F}$. Note that although this specification only includes open intervals, the σ-algebra properties mean that any σ-algebra containing these open intervals must in fact contain closed and half-closed intervals too, as well as any other sets composed of countable unions of intervals, countable intersections of intervals, and various other iterations of these processes. Let $\mathcal{B}$ be the σ-algebra generated by these intervals, known as the "Borel σ-algebra".

---

3 One might think that this argument extends further, not just to the σ-algebra generated by the propositions grasped by the agent, but to the Lebesgue completion of this σ-algebra, which adds to the generating set the collection of all subsets of sets whose probability is 0. After all, the agent's probabilities over the propositions they grasp suffice to uniquely determine the probabilities over the Lebesgue completion as well.
However, while the *original* probability function has a unique extension to the Lebesgue completion of the σ-algebra, a *distinct* probability function defined over the same initial algebra does not necessarily have a unique extension to the Lebesgue completion, because it may have a different collection of sets whose probability is 0. But if the set F of propositions grasped by the agent forms an algebra, then $\sigma(F)$ is the collection of all sets to which *all* probability functions extend uniquely, while only probability functions that assign 0 to the same subsets extend uniquely to the Lebesgue completion of $\sigma(F)$. We thank Snow Zhang for pressing this point.

4 Note that, just as in the finite case, these "worlds" aren't the fully fine-grained things that we often mean by "possible worlds". They don't fully determine everything. They just determine the truth values of the relevant propositions that the agent can directly or indirectly grasp. In general, we will use the word "world" for the atoms of the σ-algebra (and we generally assume that the σ-algebra is countably generated, which ensures that the space is partitioned into atoms).

There are several things to note about $\mathcal{B}$ and $\mathcal{C}$, and the understanding of an agent who implicitly grasps one or the other of these σ-algebras. While we normally think that the informational content of a proposition is given by the set of worlds that it contains, there is good reason to think that a proposition's informational content for an agent actually depends on the σ-algebra that the agent implicitly grasps. (See [17] for a distinct, but related set of claims.) Someone who implicitly grasps $\mathcal{B}$ is able to draw inferences that someone who implicitly grasps only $\mathcal{C}$ cannot, because $\mathcal{B}$ contains propositions that $\mathcal{C}$ does not. For instance, $\mathcal{B}$ contains the proposition ℓ consisting of all and only the points on the left half of the dial, while $\mathcal{C}$ does not. For a person with σ-algebra $\mathcal{B}$, every singleton either entails ℓ or its negation, while the person with σ-algebra $\mathcal{C}$ doesn't even implicitly grasp ℓ as a proposition, and thus can't recognize these consequences. For the one agent, the singleton carries information about which half of the dial the point is on, while for the other agent it does not.

More generally, $\mathcal{B}$ includes all the standard topological relationships among points on the dial, while $\mathcal{C}$ does not. In a sense, a person with $\mathcal{B}$ understands something about the structure of the dial that a person with $\mathcal{C}$ does not. For the person with $\mathcal{C}$, the dial consists of an unstructured continuum of points, that could just as well be scattered randomly in space, rather than being on the rim of a dial.[5]

As mentioned already, we don't necessarily need to think of the agent as directly grasping each individual proposition in their σ-algebra. It is sufficient for the agent to grasp the members of some base for this σ-algebra. Note that although $\mathcal{B}$ contains $\mathcal{C}$, $\mathcal{B}$ is countably-generated, while $\mathcal{C}$ is not. In particular, the open intervals with rational endpoints suffice to generate $\mathcal{B}$, and the countably many finite intersections of these intervals and their complements form a base for $\mathcal{B}$. We think that it is plausible that any finite physical being only directly grasps at most countably infinitely many propositions, and thus the σ-algebra of propositions they implicitly grasp must have a countable base, and thus be countably-generated.[6] Natural languages, and mathematical notation systems, enable finite beings to grasp countably many distinct propositions through their competence with finite inscriptions, but there appears to be no direct way for a finite being to grasp each element of an uncountable set.

Thus, $\mathcal{C}$ is on the one hand too complex (because it is not countably generated), and on the other hand also too unstructured (because it doesn't include topological information), to properly represent the informational content of propositions for a plausible agent. An agent with $\mathcal{C}$ would fail to grasp basic topological features of the

[5] An anonymous referee wondered whether someone whose σ-algebra is the full power set of the set of points on the dial should be thought of as grasping the topological structure of the dial, because they grasp every element of $\mathcal{B}$ and more, or as lacking this structure because, like $\mathcal{C}$, their σ-algebra is invariant under arbitrary permutations of the points, including ones that don't respect the topological structure. This may be an important question to answer for a full conceptual understanding of the structure we describe, but because of the result in footnote 6, we think the question is mostly moot.

[6] This also gives a relatively direct reason for thinking that an agent's σ-algebra will not be the full power set of Ω. Any countably-generated σ-algebra has cardinality at most the continuum, while if Ω has cardinality equal to the continuum, then its full power set must have a strictly larger cardinality. This cardinality argument for non-measurable sets sidesteps the use of symmetry considerations and Axiom of Choice constructions in the more standard Vitali argument.

dial, like which points are close to each other, and which are in a particular region of the dial. But they would also somehow have to *directly* grasp each singleton, rather than grasping most of them implicitly, by means of nested open intervals, the way someone with $\mathcal{B}$ would.

3.2 Measurability

Once we have the idea that an agent might be able to somehow grasp each *member* of a set (like Ω) without thereby even implicitly grasping every *subset* of the set, it becomes clear that not every function defined on that set is within the agent's grasp. The idea of representing some epistemologically-relevant feature by a function is that somehow, the input is sufficient to determine the output. But this naturally leads to the further thought that *information* about the input is sufficient to determine *information* about the output. To make this precise, consider some function $f: X \to Y$, and assume the σ-algebra $\mathcal{X}$ on X encodes what the agent grasps about X, while the σ-algebra $\mathcal{Y}$ on Y encodes what the agent grasps about Y. If the function f is within an agent's grasp, and if B is some graspable feature of Y (that is, $B \in \mathcal{Y}$), then the output's membership in B should be determined by some graspable feature of the input space X. Let $f^{-1}(B) = \{x \in X : f(x) \in B\}$ — the input condition determining that the output of f is in B.

> If $\mathcal{X}$ is a σ-algebra on X, and $\mathcal{Y}$ is a σ-algebra on Y, then $f: X \to Y$ is said to be $(\mathcal{X}, \mathcal{Y})$-*measurable* (or just *measurable*, if the σ-algebras are clear from context) iff for every $B \in \mathcal{Y}$, $f^{-1}(B) \in \mathcal{X}$.

We claim that every epistemologically-relevant function should be measurable with respect to the relevant σ-algebras. Non-measurable functions are functions that agents simply cannot grasp, even implicitly.

3.3 Expectations in an Infinite Setting

In Section 2.2 we defined the expected value, with respect to p, of any function $f: \Omega \to \mathbb{R}$, where Ω was finite. When Ω is infinite, not every function has an expected value. However, if p is a probability function defined on a σ-algebra $\mathcal{F}$ of Ω, and $\mathcal{B}$ is the Borel algebra on $\mathbb{R}$, then every $(\mathcal{F}, \mathcal{B})$-measurable function that has an upper bound has an expected value.[7] For the sake of completeness, we will describe how to define these expected values. Readers who aren't worried about technicalities can skip the rest of this subsection without loss of continuity.

To define expected values, we proceed in several steps. First, if f takes on only finitely many distinct values, then one of the formulas used earlier already works:

$$Exp_p(f) = \sum_{x \in \mathbb{R}} x \cdot p(\{\omega : f(\omega) = x\}).$$

[7] Unbounded measurable functions often have expected values as well, but not always. In any case, every function we take an expectation of is bounded above.

Note that $\{\omega : f(\omega) = x\}$ is $f^{-1}(\{x\})$ in the notation just defined. Since $\{x\}$ is measurable in the Borel algebra, these sets are all in $\mathcal{F}$, and thus all have probabilities given by p, so this equation is in fact well-defined. Only finitely many of these sets are non-empty, so this is a finite sum, which is thus well-defined.

If f takes on more than finitely many distinct values, but is bounded above, then we can consider the set G of all functions that are measurable, take on only finitely many values, and always take a value at least as great as the value of f. We then define:

$$Exp_p(f) = \inf_{g \in G} Exp_p(g).$$

G is non-empty, because f is bounded above, so the constant function whose value everywhere is an upper bound for the values of f is in it. Every function in G has an expectation, so this infimum exists.[8] This definition of expected value in an infinite setting is completely standard, and is equivalent to one often stated in terms of the Lebesgue integral.

3.4 Measurable Accuracy

Recall that accuracy is a function $U : \Delta(\mathcal{F}) \times \Omega \to \mathbb{R}$. That is, it takes a probability function and a world, and returns a real value measuring how accurate that probability function is in that world. While we have already specified σ-algebras $\mathcal{F}$ on Ω, and $\mathcal{B}$ on $\mathbb{R}$, we haven't yet specified a σ-algebra on $\Delta(\mathcal{F})$. Although $\Delta(\mathcal{F})$ is not a set of worlds, so subsets of it don't represent propositions, subsets of $\Delta(\mathcal{F})$ do represent features of a probability function that the agent might or might not be able to grasp. The motivating idea of measurability in this case is that, if the agent is to properly grasp a measure of accuracy, this measure must not depend on any features of probability functions that the agent is not able to grasp.

We claim that the appropriate σ-algebra $\mathcal{D}_{\mathcal{F}}$ on $\Delta(\mathcal{F})$ is the standard one, specified as follows. For any particular proposition $A \in \mathcal{F}$ and real number x, let $D_{A,x} = \{p \in \Delta(\mathcal{F}) : p(A) < x\}$. This is the set of probability functions that assign probability strictly less than x to the proposition A. It seems very reasonable to say that each of these sets can in some sense be grasped, since the agent can grasp the proposition, and can grasp the real numbers. We define $\mathcal{D}_{\mathcal{F}}$ as the σ-algebra generated by $\{D_{A,x} : A \in \mathcal{F}, x \in \mathbb{R}\}$, though any countably-generated σ-algebra containing it will work. We note that if $\mathcal{F}$ is countably-generated, then $\mathcal{D}_{\mathcal{F}}$ is also countably generated by the countably many $D_{A,x}$ where A ranges over the members of some countable base for $\mathcal{F}$, and x ranges over the rational numbers. If $\mathcal{F}$ is countably-generated, we also have that for any $p \in \Delta(\mathcal{F})$, $\{p\}$ is a member of $\mathcal{D}_{\mathcal{F}}$, as the countable intersection of all sets of probability functions whose values on the countable base are within every $1/n$ of those of p.

[8] Sometimes, this infimum might be $-\infty$ rather than a standard real number. But our assumptions will guarantee that at least some functions have finite expected value, and we are looking for the one with greatest expected value, so the fact that some have expected value $-\infty$ doesn't particularly matter.

Now that we have defined what it is for accuracy to be measurable,[9] we want to calculate expected accuracy. In Section 3.3, we showed that expectations exist for any measurable function that has an upper bound or a lower bound. So we assume that accuracy is bounded above. If we recall the definition of the omniscient probability function p_ω, where $p_\omega(A) = 1$ iff $\omega \in A$ and 0 otherwise, then it is natural to suppose that for every ω, $U(p_\omega, \omega)$ is equally maximal. As it happens, the various accuracy functions we mentioned in Section 2.3 always assigned 0 in this case, and negative numbers otherwise, as does the one example for infinite spaces that we will mention below. But even with other accuracy functions, as long as they are measurable, and bounded above, their expectations will always exist.

As before, we define $U(p)\colon \Omega \to \mathbb{R}$ by $U(p)(\omega) = U(p, \omega)$, and this time we observe that since U is measurable with respect to both input variables, and the singleton $\{p\}$ is measurable, $U(p)$ will be a measurable function of ω. Just as in the finite case, we say that U satisfies Strict Propriety iff, for any distinct $p, q \in \Delta(\mathcal{F})$, $Exp_p(U(p)) > Exp_p(U(q))$. As an example, if $\{F_i : i \in \mathbb{N}\}$ forms a countable base for $\mathcal{F}$, and the c_i are countably many positive real numbers whose sum is finite[10], then

$$U_c(p, \omega) = -\sum_{i \in \mathbb{N}} c_i (p(F_i) - p_\omega(F_i))^2$$

is a bounded (below by $-\sum_{i \in \mathbb{N}} c_i$ and above by 0), measurable (because it is a continuous function of the probabilities of a base), strictly proper[11] accuracy function.[12]

3.5 Measurable Learning

Recall that in the finite case, the set S of "signals" was a partition of subsets of Ω, and $L\colon \Omega \to S$ was the function saying what one learns in each world, sending each ω to the unique member of S containing ω. In order to state that this function is measurable, we need a σ-algebra on S. We assume that each singleton of a signal is in this σ-algebra. In order for L to be measurable, this means that each element of S must itself be a member of $\mathcal{F}$. And this seems reasonable — ordinary learning situations involve agents signals the agent could at least implicitly grasp as propositions before the event (though we will argue in Section 4.3 that this must be rejected for non-partitional updates).

[9] Technically, since U is a function from the product space $\Delta(\mathcal{F}) \times \Omega$ to $\mathbb{R}$, we need to define a σ-algebra over $\Delta(\mathcal{F}) \times \Omega$, rather than just a σ-algebra $\mathcal{D}_\mathcal{F}$ over $\Delta(\mathcal{F})$ and a σ-algebra $\mathcal{F}$ over Ω. We define $\mathcal{D}_\mathcal{F} \otimes \mathcal{F}$ to be $\sigma(\{D \times A : D \in \mathcal{D}_\mathcal{F}, A \in \mathcal{F}\})$. This is the σ-algebra generated by the product of the σ-algebras over the individual sets.

[10] The sum of the c_i being finite is not just sufficient, but necessary for Strict Propriety of this function [24]. We thank an anonymous reviewer for pointing us to this paper.

[11] To see that it is strictly proper, we again calculate that $Exp_p(U_c(p)) - Exp_p(U_c(q)) = \sum_{i \in \mathbb{N}} c_i (p(F_i) - q(F_i))^2$, which is minimized iff each $p(F_i) = q(F_i)$. But since the F_i form a base for $\mathcal{F}$, this means that p and q are in fact identical.

[12] Note that [31] shows that there are no strictly proper accuracy functions defined over *all* functions from $\mathcal{F}$ to $\mathbb{R}$, or that are strictly proper for *finitely additive* "probability" functions. But we only require the accuracy function to be defined and Strictly Proper for the countably additive probability functions.

Dubra & Echenique [9] show that a problem arises if we use just the σ-algebra generated by the singletons to measure the informational content of the signals. They note that intuitively, a finer partition should convey more information than a coarser partition. But if the information content of a signal is given by the σ-algebra generated by the singletons, then the finest partition into points generates the countable/co-countable σ-algebra, which doesn't contain the information conveyed by an extremely coarse partition into the left half and right half.[13] As we pointed out above, the countable/co-countable σ-algebra doesn't contain topological information.

We can say that partition S is *finer* than S' (and S' is *coarser* than S) iff every member of S' is a union of members of S. That is, S draws all the distinctions drawn by S', and possibly more. The idea suggested by Dubra and Echenique is that the σ-algebra associated with a finer partition should always contain any information contained in the σ-algebra associated with a coarser partition. Since any member of $\mathcal{F}$ that is a union of members of S could be part of a coarser partition, this means that the σ-algebra associated with S should contain any such proposition. That is, rather than the σ-algebra on S being the *smallest* σ-algebra containing each singleton, we use the *largest* σ-algebra that makes L measurable. (This solution is endorsed by [38].) We refer to this σ-algebra as $\mathcal{S}$, and call it the sub-σ-algebra of $\mathcal{F}$ that is *informed* by S.

Although there exist partitions S such that $\mathcal{S}$ is not countably-generated[14], we think that these are not epistemologically relevant. In order for an agent to grasp a learning situation, it is not sufficient that they grasp each possible signal that they learn — there must be some countable set of propositions that they grasp that suffice to distinguish these signals. Thus, we assume that the σ-algebra $\mathcal{S}$ on S is the informed sub-σ-algebra of $\mathcal{F}$, and that the set of signals S is such that $\mathcal{S}$ is countably-generated.

3.6 Measurable Update Plans and Regular Conditional Probabilities

An update plan is a function $P: S \to \Delta(\mathcal{F})$. In the previous two sections, we came up with a σ-algebra $\mathcal{S}$ on S, and a σ-algebra $\mathcal{D}_{\mathcal{F}}$ on $\Delta(\mathcal{F})$. We argued that $\mathcal{S}$ consists of all subsets of S whose union is in $\mathcal{F}$, and that $\mathcal{D}_{\mathcal{F}}$ is generated by the sets of probability functions defined by specifying whether a particular proposition A has probability below some real threshold x. Using these σ-algebras, an update plan is measurable iff every fact about whether or not some particular proposition A has probability below some threshold x is determined by some set of signals that, together, form a proposition that the agent at least implicitly grasps.

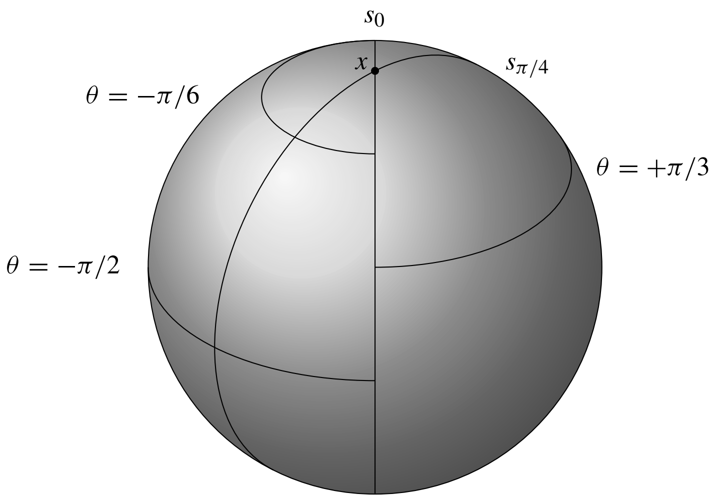
Just as in the finite case, we define a function $U(P): \Omega \to \mathbb{R}$ by letting $U(P)(\omega) = U(P(L(\omega)), \omega)$. Because U, P, and L are all measurable, so is this function. Thus, we can calculate $Exp_p(U(P))$, and ask for an update plan P that maximizes it. Nielsen [29] proves a generalization of the result of Greaves and Wallace [16] (that also generalizes the results of Easwaran [10]). Just as in the finite case, when $s \in S$

[13] Seidenfeld et al. [35] find a related problem that occurs when we use the countable/co-countable σ-algebra on S, namely that the optimal update on this σ-algebra almost always ignores the signal, leaving the probability function unchanged.

[14] Consider $\Omega = [0, 1]$, and let each member of S consist of a number in this interval, together with all other numbers in this interval that differ from it by a rational number. [37] (p. 242) shows that there is no countably-generated algebra that has these sets as its atoms.

with $p(s) > 0$, we can define $p_s(A) = p(A \cap s)/p(s)$, and show that for any such s, $P_s = p_s$ whenever P maximizes expected accuracy.

But in the infinite case, it is quite possible to have $p(s) = 0$ for a larger class of signals s. This can even happen for all s. To see this, let Ω be the set of points on the two-dimensional surface of a sphere in $\mathbb{R}^3$. Let $\mathcal{F}$ be the Borel σ-algebra on Ω, and let p be the uniform probability. Pick some point x on the surface of the sphere, and let s_0 be a particular great circle through this point. For $0 < \varphi < \pi$, let s_φ be the set of all points other than x and its antipode that lie on the great circle through x that lies at a (directed) angle of φ from s_0. Let $S = \{s_\varphi : 0 \leq \varphi < \pi\}$. These signals correspond (roughly[15]) to measurement of the longitude of a point, if x is the north pole (with the north and south poles being arbitrarily defined to have longitude 0). Then in this case, $p(s) = 0$ for all $s \in S$. As a result, p_s is undefined for all s, so the result that $P_s = p_s$ when $p(s) > 0$ gives us no information about how to maximize expected accuracy when learning which line of longitude a point is on.



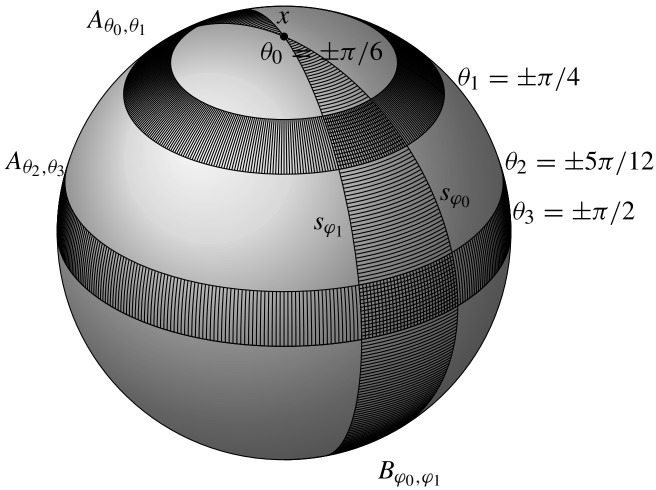
However, Nielsen demonstrates another feature that is in fact sufficient to characterize update by maximizing expected accuracy. Recall that $P^A : S \to \mathbb{R}$ is given by $P^A(s) = P_s(A)$, mapping any signal s to the probability the update plan would assign to A after receiving it. For any subset T of S, let $1_T : S \to \mathbb{R}$ be the indicator function for T defined by $1_T(s) = 1$ if $s \in T$ and 0 otherwise. Note that $\bigcup T$ is the set of worlds in which one receives a signal that is a member of T. We say that an update plan P is a *regular conditional probability* for p on $\mathcal{S}$ iff

$$\text{For every } A \in \mathcal{F} \text{ and } T \in \mathcal{S},\ Exp_p\Big(1_T(L(\omega)) \cdot P^A(L(\omega))\Big) = p(A \cap (\bigcup T)).$$

[15] In standard geography, the lines of longitude are semicircles, both west and east of the 0 meridian. Here, they are full great circles, so that the semicircle that is at $\pi/3$ east is actually part of our great circle that is $2\pi/3$ west. This is why φ only goes up to π, rather than 2π.

Roughly, the average posterior probability of A in the T worlds is the prior probability of A in the T worlds. Nielsen shows that an update plan maximizes expected accuracy iff it is a regular conditional probability.[16]

To illustrate some of the implications of this, let us continue the above example. Ω is the surface of a sphere, p is the uniform probability distribution, and S is the set of lines of longitude. We can represent each ω as (φ, θ), where $0 \leq \varphi < \pi$ indicates which longitude s_φ it is in, and $-\pi \leq \theta < \pi$ indicates the directed distance from the North Pole to the point along the great circle.[17] Then $\mathcal{S}$ will contain all measurable sets composed of lines of longitude, and in particular will contain any event $B_{\varphi_0,\varphi_1} = \{(\varphi, \theta) : \varphi_0 < \varphi < \varphi_1\}$. For $0 < \theta_0 < \theta_1 < \pi$, let $A_{\theta_0,\theta_1} = \{(\varphi, \theta) : \theta_0 < |\theta| < \theta_1\}$ — this is the event that a point has latitude between θ_0 and θ_1. Some calculus shows us that $p(A_{\theta_0,\theta_1}) = (\cos\theta_0 - \cos\theta_1)/2$. Based on the rotational symmetry of the setup, we can show that, in this case, P is a regular conditional probability for p on S iff for almost all[18] s, $P_s(A_{\theta_0,\theta_1}) = p(A_{\theta_0,\theta_1})$ for all A_{θ_0,θ_1} (this suffices to determine P_s since the A_{θ_0,θ_1} form a base for the Borel σ-algebra of subsets of s). In particular, this is a non-uniform probability on the great circle s, being concentrated towards the equator.



Although $\theta_1 - \theta_0 = \theta_3 - \theta_2$, so A_{θ_0,θ_1} and A_{θ_2,θ_3} intersect each s_φ in the same length, the update that maximizes expected accuracy on the s_φ is proportional to the *areas* rather than the lengths.

16 Note that in general, regular conditional probabilities are not unique. If P is a regular conditional probability for p on $\mathcal{S}$, and $P' : S \to \Delta(\mathcal{F})$ differs on an $\mathcal{S}$-measurable set whose union has probability 0, then P' is also a regular conditional probability for p on $\mathcal{S}$.

17 This is an even more unorthodox measure of "latitude", but it corresponds to the same lines as ordinary latitudes. The South Pole has "latitude" $\pm\pi$, the North Pole has "latitude" 0, the equator has "latitude" $\pm\pi/2$, and points in the Western Hemisphere have negative "latitude" while points in the Eastern Hemisphere have positive "latitude".

18 That is, the set of all s for which this is not true together have probability 0. It is clear that changing the values of P on some set with probability 0 doesn't change the expected accuracy of P, and so this sort of "almost all" condition is the most that can be stated on the basis of expected accuracy.

In general, it is known that such regular conditional probabilities exist whenever $\mathcal{F}$ is isomorphic to the Borel σ-algebra on some Borel subset of $\mathbb{R}^n$, as it is in every one of our examples.[19] This is the orthodox theory of conditional distributions in mathematical probability [25]. There have been various objections to the use of regular conditional probabilities for update plans. Many of these objections arise only when $\mathcal{F}$ and $\mathcal{S}$ fail to be countably-generated, or when $\mathcal{S}$ fails to be the sub-σ-algebra of $\mathcal{F}$ informed by S.[20] Since we assume that $\mathcal{F}$ and $\mathcal{S}$ are countably generated, and that $\mathcal{S}$ is the sub-σ-algebra informed by S, these worries do not arise.

But probably the most central objection to regular conditional probabilities is often referred to as the "Borel-Kolmogorov Paradox". To illustrate this, return once again to the example where Ω is the sphere, p is the uniform probability, and S is the set of lines of longitude through some given point x. Let x' be some point distinct from x and its antipode, and let S' be the set of lines of longitude through that point. There is exactly one s that is a member of both S and S', namely the great circle through x and x'.[21] As we noted, if P is a regular conditional probability for S and P' is a regular conditional probability for S', then P_s is non-uniform, concentrated towards the equator, while P'_s is non-uniform, concentrated towards *its* equator. Since these equators are different, these are different non-uniform distributions, even though the same signal was received in the two cases. The set of signals the signal was drawn from made a difference to how the agent should update on the signal.

Kadane et al. [23] (pp. 224-5) say this is "unacceptable from the point of view of the statistician who, when given information that [the event] has occurred, must determine the conditional distribution." They think that any general class of update policies should be "coherent", meaning that receiving the same signal should result in the same update, regardless of what learning situation it was received in.[22] Based on our result, this would require giving up the idea that updates should maximize expected accuracy. Rescorla [32] argues that instead, we should see this as a kind of "Frege puzzle", where one and the same proposition is presented in two different "guises", so that one can have different epistemological reactions to the same signal.

But in this setting, we think that it is misleading to treat the signal itself, considered just as the set of worlds, as sufficient to determine the content that is updated on. Rather, as mentioned in Section 3.1, the content is partially determined by the σ-algebra this signal is embedded in. Just as a singleton has different content for an agent who implicitly grasps the Borel σ-algebra (for whom it implies facts about what intervals the point is in) as opposed to an agent who only implicitly grasps the countable/co-countable σ-algebra (for whom it does not), the same great circle has different content for agents who learn it as one of the longitudes through a different pair of poles. The wedges that are composed of intervals of lines of longitude with

[19] Both algebras being countably generated is not sufficient. A counterexample described in many standard textbooks has $\Omega = [0, 1]$, $\mathcal{F}$ the σ-algebra generated by the Borel sets and one additional subset, while S is the Borel sets. We have not yet identified a fully philosophical motivation for conditions that are sufficient for the existence of regular conditional probabilities, but every natural example satisfies them.

[20] For instance, see many of the challenges raised by [35].

[21] It's to make this true that we chose the non-standard account of "longitudes" as great circles rather than great semicircles.

[22] For more on regular conditional probabilities and coherent conditional probabilities, see [12].

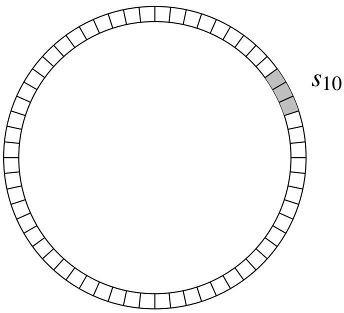
respect to one pair of poles are distinct from the wedges composed of intervals of lines of longitude with respect to a different pair of poles, and no wedge from one set is even a subset of any wedge from the other set. And thus we think that, although the same *signal* is received, the *content* is different, and it is this different content that results in a different update.

In any case, when we return to a related example in Section 5.3, we show that this paradox often disappears in the infinite, non-partitional setting.

4 Non-partitional Learning

In addition to the setting where Ω is infinite, the other generalization we make of the Greaves and Wallace result is to the setting where S does not consist of a partition whose elements are members of $\mathcal{F}$. Many authors in this literature consider S as some collection of overlapping elements of $\mathcal{F}$, but we will eventually show that the identification of elements of S with elements of $\mathcal{F}$ is inessential to maximizing expected accuracy — all that matters is the "likelihoods" of these elements (to be discussed in Section 4.3). Thus, we will eventually treat the elements of S as un-analyzed "signals" (though we will consider some interpretations of how to think of them in Section 4.4). But at first, we will follow others and identify elements of S with elements of $\mathcal{F}$.

In this section, we return to the assumption that Ω is finite but allow S to not be a partition of Ω. To motivate this setting, consider Williamson's [39] case of the unmarked clock. In this case, the agent is about to look at an "irritatingly austere" clock that has a minute hand, but no other markings on its face. The agent knows that the hand of the clock will point at one of the 60 precise spots corresponding to the different minutes, but also knows that their visual system is not precise enough to tell precisely which of these spots the hand is pointing at.



To model this, we let $\Omega = \{\omega_0, \omega_1, \ldots, \omega_{59}, \omega_{60} = \omega_0\}$, so that each world just specifies which precise position the minute hand is pointing at. (We let the topmost position be named both "ω_{60}" and "ω_0" for convenience.) Williamson says that we can represent limitations of the agent's vision by saying that, in each world, when the agent looks at the clock, their evidence will consist of the set containing that world and the two adjacent worlds. That is, in ω_i, their evidence will be the proposition $s_i = \{\omega_{i-1}, \omega_i, \omega_{i+1}\}$. Importantly, although this proposition is in fact the agent's evidence in ω_i, it doesn't entail *that* it is their evidence, since it is not in fact their evidence in two of those worlds. Any situation in which what one learns is a proposition that fails to entail that one learns that very proposition is non-partitional. We will eventually suggest that, in non-partitional cases, it is more helpful to think of signals abstractly than to try to characterize the learning situation with propositions. We will start with analyses from the literature that do identify signals with propositions, but readers can skip ahead to our analysis in Section 4.3.

4.1 Schoenfield's Analysis

Schoenfield [33] treats the case of the unmarked clock much like how we have treated the original [16] framework above, with the one exception that the members of S are overlapping sets. The learning function L thus can't assign each ω to the *unique* member of S where it is true, but instead just has $L(\omega_i) = s_i$. Let us define $L^{-1}(s) = \{\omega : L(\omega) = s\}$ as the set of worlds in which the agent receives s as their signal. In the partitional setting, $L^{-1}(s) = s$, but in the non-partitional setting, this is no longer generally true. In particular, for the unmarked clock, $L^{-1}(\{\omega_{i-1}, \omega_i, \omega_{i+1}\}) = \{\omega_i\}$.

In the original framework, we maximized expected accuracy by maximizing $Exp_{p_s}(U(P_s))$. But the signal s plays two separate roles in this framework. In particular, in "p_s", it represents the set of worlds in which the agent receives s as their signal, while in "P_s", it represents the signal that the agent uses to drive their update. Since s is no longer the set of worlds in which the agent receives s as their signal, these roles are not played by the same entity. Instead, we should replace the former "s" by "$L^{-1}(s)$", yielding the result that the update plan that maximizes expected accuracy will be the one that maximizes $Exp_{p_{L^{-1}(s)}}(U(P_s))$. This will occur whenever $P_s = p_{L^{-1}(s)}$. That is, the plan that maximizes expected accuracy doesn't correspond to conditionalizing on one's evidence, but instead corresponds to conditionalizing on the proposition *that* one received this particular piece of evidence. This rule has been called "meta-conditionalization" or "auto-epistemic-conditionalization" [7, 18].

In the case of the unmarked clock, this gives potentially unexpected, or even unwanted, results. In particular, each world yields its own distinct piece of evidence, and so the proposition that is conditionalized on is a singleton. Conditionalizing on a singleton $\{\omega\}$ yields the omniscient probability function p_ω that assigns the value 1 to every proposition true in that world, and 0 to every proposition false in that world. That is, despite the agent's imperfect vision, which prevents them from knowing precisely where the hand of the clock is pointing, Schoenfield's recommendation is that they plan to update by becoming certain of precisely where the hand of the clock is pointing.

For other cases in which the same evidence might be had in multiple worlds, the result is not quite as strong. The agent won't always be recommended to use an update plan that is certain precisely which world is actual. However, they will still be recommended to use an update plan that makes them certain of $L^{-1}(s)$, which will often be a proper subset of s if each s is a possible evidence proposition and these evidence propositions overlap.

This is the result of asking that the update plan be a function of the signal, when the signal is a deterministic function of the world. However, Schoenfield's official formalism is, like that of Greaves and Wallace, that the update is a function of the world directly, rather than having the signal as an intermediary. Greaves and Wallace don't ask for the function from worlds to probability functions that maximizes expected accuracy (which would always be the one that yields, in each world, the probability function that takes the value 1 for each proposition true at that world and 0 for each proposition false at that world), but instead for the *available* update plan that maximizes expected accuracy. They say that an update plan is available iff it assigns the same probability function for any two worlds that appear in the same possible evidence proposition, which is equivalent (in the partitional setting) to saying that a function is available iff it assigns the same probability function for any two worlds in which the agent receives the same evidence. But Schoenfield notes that these are different in the non-partitional setting. In particular, in the case of the unmarked clock, requiring the update plan to assign the same probability function in any two worlds that appear in the same possible evidence proposition ends up requiring the update plan to assign the same probability function in *all* worlds, which rules out intuitively possible updates, just as Schoenfield's framework rules in intuitively impossible updates.

4.2 Four Recent Analyses

Carr, Gallow, Isaacs & Russell, and Schultheis [6, 14, 19, 34] have given alternative analyses of this case, and the general problem of updating by maximizing expected accuracy in a finite non-partitional setting. These papers give substantially different philosophical interpretations of the situation. We believe that much of the accounts of Gallow and Isaacs & Russell can be described with the formalism we develop, interpreted somewhat differently. We will briefly describe the interpretations of each, and then give the details of the formalism, and the central results they share. We think this formalism naturally admits of multiple interpretations, including some we propose that differ significantly from those of these other authors.

A slightly different account is given by [6], who evaluates updates in a non-partitional setting not by asking that the agent have a plan *before* receiving the signal that maximizes expected accuracy, but instead by asking *after* the agent receives the signal, that their update be endorsed by a "nonconsequentialist decision theory", which evaluates the update in all the worlds compatible with their signal (whether or not they would have updated the same way in those worlds). This evaluation is different enough from ours, and from the others that we consider, that we skip a fuller description here. (It does have some surprising similarities to the proposal made by Konek [26] for an apparently different type of update.) The result she endorses always recommends that agents conditionalize on the proposition that is their evidence. We will see that this is at least sometimes distinct from the result endorsed by the others (including Konek), which we end up endorsing too.

Gallow and Schultheis both follow Schoenfield and Williamson in saying that the world determines what one's evidence is, and that the evidence is a proposition consisting of a set of worlds. They both keep the idea that an update plan is a function from evidence propositions to probability functions. However, they both allow that the evidence proposition that one *updates* on might not be one's *actual* evidence. Gallow suggests that the agent might "misfire" and accidentally update on a different evidence proposition than the one they actually have as evidence. Schultheis thinks that while the way to evaluate the best update plan to "follow" is that described by Schoenfield, we should instead evaluate what is the best update plan to "adopt". If one "adopts" an update plan, one might "misapply" it, because one "thinks that" the evidence is some proposition other than the one that it actually is. In each case, it is natural to continue to think of some member of S as one's "evidence", and to think of P as a function from S to probability functions, but to allow that the member of S one updates on might not be the member of S that is one's evidence.

Both Gallow and Schultheis instead say there is a probability function, determined by the world one is in, saying how likely one is to update on each potential signal in S. Our formalism will agree with this, except that we will dispense with the assumption that each signal is a proposition that could be one's evidence in some world. (As we will see, the interpretation of this probability function is a bit tricky. Schultheis suggests this raises difficulties for finding the optimal update plan, though she does show that some plans are suboptimal, while we follow Gallow in attempting to identify an optimal update plan.)

Isaacs and Russell give up on the idea that there is any need for the concept of "evidence". Instead, they define a "stochastic guess model", where the update plan determines one's updated probability function on the basis of one's "guess", which is not uniquely determined by the world one is in. They think of these guesses as guesses about which world one is in, and they think that one can have any plan as a function from worlds to probabilities but that one will apply it to one's guess, rather than to the actual world. The world ω does not uniquely determine one's guess, but instead yields a probability function over one's possible guesses. Their "guesses" will again correspond to our signals, but where they make the assumption that $S = \Omega$ (i.e., each guess is a guess about which world is actual) we allow that signals might be drawn from a distinct set, though we recover their results if we make their assumptions.

4.3 Non-Partitional Signals

Although some of these authors think of the members of S as propositions that one thinks are one's evidence, and others think of them as guesses about worlds, we will refer to them neutrally as "signals". While they might have propositional content in some sense, we think that what is important about them is that they capture what the agent is able to respond to in the update situation in question. Because the agent *is* in fact able to respond to them, it is appropriate to evaluate potential update rules considered as functions from them to probability functions.

Recall that a "world" here is not a totally fine-grained entity that determines every fact. When we assume that the set of "worlds" is finite, we already assume that these "worlds" are more coarse-grained, so that this sort of probability function could conceivably make sense. Even in the infinite setting, as we mentioned in footnote 4, these "worlds" just represent the limits of what the agent distinguishes, and not the precise specification of everything there is. In particular, we don't assume that the world specifies the agent's credence function or update — we don't make use of the agent's credences about their own credences.[23]

[23] Some authors in this literature, like [5, 6, 14], are particularly interested in understanding uncertainty about how one updated, but this is not a focus of ours.

One might think that if the agent distinguishes the different signals they might receive, then the worlds should at least be finely-enough divided to determine the signal the agent receives. But non-partitional learning motivates rejecting this assumption. The set of worlds in which one might receive signal s may overlap with the set of worlds in which one might receive signal s'. L is no longer a function from worlds, as conceived by the agent, to signals. Instead, we think of $L\colon \Omega \to \Delta(\mathcal{S})$ (where $\mathcal{S}$ is the σ-algebra consisting of all subsets of S, since S is finite), so that each world is associated with some *probability function* over the possible signals S, rather than with a single signal. We now use "L" to stand for the "likelihood" of each particular world producing each particular signal, rather than "learning". (This roughly corresponds to Gallow's probability of "misfire", Schultheis's probability of "following" a rule distinct from the one the agent adopts, and Isaacs and Russell's "stochastic guess".)

There is a question of how to interpret this function L. We will treat it as an objective probability whose expectations are objectively relevant to which plan for the agent maximizes expected accuracy. (As we will see, there are questions about what "expected accuracy" means on the relevant probability function, but we think these questions are just as relevant for Gallow, and for Isaacs and Russell, as for us.) Just as we often write "P_s" rather than $P(s)$ for the probability function P tells the agent to adopt on receiving signal s, we will often write "L_ω" rather than $L(\omega)$, again in order to avoid awkward multiple sets of parentheses.

Gallow and Schultheis both identify signals s with the "evidence" the agent receives, which is itself identified with a set of worlds, even though the agent doesn't update on s in all those worlds. Isaacs and Russell identify s with the world one "guesses" is actual, even though one may or may not make that guess in that world or in various others. These are possible interpretations of our framework, but we think some other possible interpretations are equally good.

One possibility is on analogy with an idea of Frank Jackson [20, 21]. He argues that there is a kind of knowledge one can gain from experience that is distinct from anything phrased in terms of physical concepts. We might treat the signal as bearing this kind of phenomenal knowledge. An anonymous referee has suggested that a less mysterious way to accomplish the same thing is to treat the signal as the neural state the agent is in as a result of their perceptual experience, which again, is not usually conceptualized by the agent.

Another possibility follows some ideas from the literature on "knowledge how" and "knowledge that". Stanley & Williamson [36] argue that "knowledge how" is in fact a species of "knowledge that", but that it involves the proposition being known under a special "practical mode of presentation". That is, they claim that the sentence "Hannah knows how to ride a bicycle" expresses the same knowledge that "Hannah knows that *that* is a way for her to ride a bicycle", where the "*that*" picks out some way of riding a bicycle in a way that she can't express in any way other than by ostension. They analogize this to special first-personal modes of presentation. We think there is a similar possibility in cases like the un-marked clock — after looking at the clock, one knows that *that* is how the clock looks, even if one doesn't know whether the clock looks like it is between 1:05 and 1:07 or looks like it is between 1:06 and 1:08 or looks like it is between 1:07 and 1:09. All that is relevant for the proposition that is known in Stanley and Williamson's analysis of "knowledge how" is that it is available to be turned into action, and all that is relevant for the signal in the present context is that it is available to be turned into an update of one's credences.

In any case, we say the σ-algebra $\mathcal{F}$ consists of the propositions the agent grasps in the "worldly" way, and the σ-algebra $\mathcal{S}$ consists of the propositions the agent grasps in the "experiential" way. It is not essential to our formalism or our philosophical argument whether this truly involves internalistic non-conceptual phenomenology or externalistic reaction to features of the world that one doesn't conceptualize — what is essential is that $\mathcal{S}$ represents the sorts of propositions that the agent has the implicit capacity to respond to, even if their only grasp of these propositions is through some sort of experiential or practical demonstrative.

Probability functions in $\Delta(\mathcal{F})$ (such as the agent's prior p) express potential ways for the agent to have uncertainty about propositions presented in a worldly way, and probability functions in $\Delta(\mathcal{S})$ (such as any of the L_ω) express chancy ways for the agent's update to depend on what the agent grasps in a worldly way. When Ω and S are both finite, as we currently assume, we can put together p and L to get something that is formally a probability function $p \times L$ over $\Omega \times S$, with $(p \times L)(\{(\omega, s)\} = p(\{\omega\}) \cdot L_\omega(\{s\})$. But because this probability function involves both Ω and S, it applies to many propositions that aren't grasped in either the worldly or the experiential way, but only in some sort of mixed way.

This formalism is compatible with the interpretations given by Gallow, Isaacs and Russell, and Schultheis. For Gallow and for Schultheis, S corresponds to the set of possible pieces of evidence one might have, and L encodes the probability of "misfiring" or "misapplying" when implementing one's update plan. They combine these with the prior p to generate a probability over the pairs (ω, s), representing the situation when ω is the actual world, and s is the evidence the agent (perhaps mistakenly) uses in their update.[24] For Isaacs and Russell, S corresponds to the set of possible worlds, just as Ω does, but represents the world qua "guess" rather than qua actuality. They work directly with a probability distribution over $\Omega \times S$ (or $\Omega \times \Omega$ in their notation), while we work with p and L separately and combine them to yield this product probability. But as we will note in Section 4.5, our framework also allows the discussion of cases where S corresponds neither to some subset of $\mathcal{F}$ nor to Ω. If the formalism of any of these authors admits of a philosophically applicable interpretation, then ours admits of that interpretation too (and can generalize it to infinite cases) though we suggest the possibility of more general interpretations.

[24] More specifically, they assume that any s is itself a subset of Ω, and that there is some set of worlds where it is in fact your evidence. Gallow refers to this set as "$\mathbf{T}s$", and uses "$\mathbf{U}s$" to denote what we represent as the set of ordered pairs whose second element is s. Schultheis effectively works directly on the finer-grained space of pairs (ω, s) and has an "extension function" that does the work of our $p \times L$, though she only explicitly considers functions that have a particular symmetry, and only shows that conditionalization is non-optimal, rather than computing the optimal update plan. She has principled reasons for worrying about the ability of an agent to engage in an optimal update plan. We try to address these worries in Section 4.5, but acknowledge that they may be foundational obstacles to any program like ours, or that of Gallow, or of Isaacs and Russell.

4.4 The Result

Just as in the finite partitional case, as described in Section 2, Ω is a finite set of worlds, and $\mathcal{F}$ is the full power set of Ω. The agent's prior probability function is $p \colon \mathcal{F} \to \mathbb{R}$. There is a finite set S of possible signals that the agent might receive, but we no longer interpret these signals as elements of $\mathcal{F}$. Letting $\Delta(\mathcal{F})$ represent the set of all probability functions on $\mathcal{F}$, we think of an update plan as a function $P \colon S \to \Delta(\mathcal{F})$, that yields a probability function for any possible input signal. Just as before, accuracy is given by an epistemic utility function $U \colon \Delta(\mathcal{F}) \times \Omega \to \mathbb{R}$, and we assume that this is strictly proper, so that $Exp_p(U(p)) > Exp_p(U(q))$ whenever p and q are distinct probability functions over Ω. Note that U only applies to worldly probability functions over Ω, not probability functions over S or $\Omega \times S$, though expectations of U can be calculated over any probability space whose members can be fed into U somehow or other.

In our discussion of the partitional setting in Section 2.4, we assumed that $L \colon \Omega \to S$ uniquely determined the signal one would receive in every world, and defined $U(P) \colon \Omega \to \mathbb{R}$ as the function that took world ω to the accuracy it would yield if applied in world ω, namely $U(P(L(\omega)), \omega)$. We then chose P to maximize the expectation $Exp_p(U(P))$. But in the new setting, the world doesn't uniquely determine the signal. Instead, we have $L \colon \Omega \to \Delta(\mathcal{S})$ giving a probability distribution over signals. Furthermore, we must instead define $U(P) \colon \Omega \times S \to \mathbb{R}$ by the function that takes the pair (ω, s) to the accuracy P would yield if applied in world ω with signal s, namely $U(P_s, \omega)$. Expected accuracy for such an update plan must then be defined relative to the probability function $p \times L$, rather than just p.

We will again define an auxiliary probability function $p_s \colon \mathcal{F} \to \mathbb{R}$, but this time it will have the value

$$p_s(A) = \frac{\sum_{\omega \in A} p(\{\omega\}) \cdot L_\omega(s)}{\sum_{\omega \in \Omega} p(\{\omega\}) \cdot L_\omega(s)}.$$

If we were to take $p \times L$ to be the agent's prior over the set $\Omega \times S$, then this is the probability function that would result from conditioning it on the set of pairs whose second element is s. The numerator is the sum of the probabilities of the pairs consisting of a world where A is true together with signal s, while the denominator is the sum of the probabilities of all pairs with signal s. Note that both p and p_s are probability functions over worlds, not over signals, even though p_s is defined in terms of $p \times L$, which involves both worlds and signals. The one mathematical fact about this auxiliary probability function that is relevant for our calculation (other than the fact that it is in fact a probability function) will be that $(p \times L)((\omega, s)) = p(\{\omega\}) \cdot L_\omega(s) = (p \times L)(\Omega \times \{s\}) \cdot p_s(\{\omega\})$.

Now we can chase some definitions again.

$$\begin{aligned} Exp_{p \times L}(U(P)) &= \sum_{\omega \in \Omega} \sum_{s \in S} (p \times L)((\omega, s)) U(P_s, \omega) \\ &= \sum_{s \in S} \sum_{\omega \in \Omega} (p \times L)(\Omega \times \{s\}) \cdot p_s(\{\omega\}) U(P_s, \omega) \end{aligned}$$

$$= \sum_{s \in S} (p \times L)(\Omega \times \{s\}) \sum_{\omega \in \Omega} p_s(\{\omega\}) U(P_s, \omega)$$
$$= \sum_{s \in S} (p \times L)(\Omega \times \{s\}) Exp_{p_s}(U(P_s))$$

The first step just switched the order of the summation and applied the previous result about our auxiliary probability function. The second step factored out the terms that depend only on s and not ω. The third step applied the definition of expectation over this auxiliary probability function.

The last line reveals the important fact. Each probability function P_s resulting from update on a particular signal s contributes a separate term to the overall expected accuracy of P. We can maximize the expected accuracy of P by maximizing the contribution of each term P_s. Because each term contributes through a non-negative multiple of $Exp_{p_s}(U(P_s))$, and because P_s and p_s are both probability functions over Ω, for which we can apply Strict Propriety, we see that this is maximized iff $P_s = p_s$ for all s.

This resulting update policy is just as if the agent actually had the probability distribution $p \times L$ over $\Omega \times S$, and conditionalized on having received the signal s.[25] As Schultheis notes, one problem with using conditionalization for these sorts of cases is that one's evidence doesn't entail *that* it is one's evidence, and so one shouldn't be able to rationally become certain that it is one's evidence. One might think that this also means that one isn't able, on the basis of some signal, to become certain *that* it is the signal one used. However, for Schultheis, the fact that something is one's *evidence* is a worldly proposition. But in our setting, the thing one becomes certain of is in the *experiential* mode of presentation, not the worldly mode of presentation. It is as if one is certain of what signal one used, not as if one is certain of what evidence one actually has. Stanley & Williamson[36] say that, by riding a bike, an agent can know that *that* is a way for them to ride a bike, even without being able to put "*that*" into words. Similarly, we think that by looking at a clock at 1:06, an agent can know that *that* is how the clock looks to them without "that" being equivalent to any proposition expressible in terms of time.

But formally, propositions about the signal, even in this experiential mode of presentation, are not part of the algebra $\mathcal{F}$ that we represent the agent as having credences over, so our results don't actually imply that the optimal update will result in certainty about signals. (It is only *as if* one had this certainty.) We don't interpret L as the agent's uncertainty about what signal they would update on in each world, but rather as some sort of objective probability of the agent activating the part of their update plan driven by signal s in a given world. On this interpretation, $p \times L$ is a mixture of subjective and objective probabilities, so it is not fully clear what the implications are

[25] This result is formally identical to one that Andrew Bacon [1] derives for update via conditionalization on vague propositions. The idea is described in his Chapter 6, and almost precisely the same formula that we give for p_s occurs in section 8.1, though with a different interpretation, and without a justification using expected accuracy.

of an update maximizing expected accuracy with respect to this probability function. But we think that it is still suggestive, the way that the mixture of credal dominance and chance expectations is in [30]. In any case, the same interpretative difficulties arise for the "guess distribution" of Isaacs and Russell and the "misfire" distribution of Gallow.

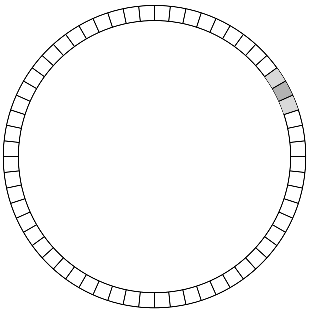
4.5 Examples

Returning to the case of the unmarked clock, let us assume again that $\Omega = \{\omega_0, \omega_1, \ldots, \omega_{59}, \omega_{60} = \omega_0\}$, and let $S = \{s_0, s_1, \ldots, s_{59}, s_{60} = s_0\}$. If, for all i, we have $L_{\omega_i}(\{s_{i-1}\}) = L_{\omega_i}(\{s_i\}) = L_{\omega_i}(\{s_{i+1}\}) = 1/3$, then regardless of one's prior p, the plan that maximizes expected accuracy will be the same as planning to conditionalize p on $\{\omega_{i-1}, \omega_i, \omega_{i+1}\}$ if one receives the signal s_i. This might lead one to identify signal s_i with this worldly proposition.

However, note that the proposition that s_i is one's signal corresponds to the set $\{(\omega_{i-1}, s_i), (\omega_i, s_i), (\omega_{i+1}, s_i)\}$ in the combined algebra, which is distinct from the worldly set $\{\omega_{i-1}, \omega_i, \omega_{i+1}\}$, which corresponds in the combined algebra to

$$\{(\omega_{i-1}, s_{i-2}), (\omega_{i-1}, s_{i-1}), (\omega_{i-1}, s_i), (\omega_i, s_{i-1}), (\omega_i, s_i),$$
$$(\omega_i, s_{i+1}), (\omega_{i+1}, s_i), (\omega_{i+1}, s_{i+1}), (\omega_{i+1}, s_{i+2})\}.$$

Conditioning on them happens to yield the same result for worldly propositions, but this is a misleading artifact of the fact that for every signal, every world that gives it a non-zero probability gives it the same probability. (This fact is a generalization of Theorem 4 of Isaacs and Russell.)



To see the distinction, consider another case, where $L_{\omega_i}(\{s_{i-1}\}) = L_{\omega_i}(\{s_{i+1}\}) = 1/4$ and $L_{\omega_i}(\{s_i\}) = 1/2$. This represents a situation where each signal is possible in three worldly situations, but is more likely in the middle one of the three. In many ways, this actually seems like a more natural description of the situation with a real unmarked clock. In this case, $p_{s_i}(\{\omega_{i-1}\})$, $p_{s_i}(\{\omega_i\})$, and $p_{s_i}(\{\omega_{i+1}\})$ are proportional to $(1/4)p(\{\omega_{i-1}\})$, $(1/2)p(\{\omega_i\})$, and $(1/4)p(\{\omega_{i+1}\})$ respectively, but they must be renormalized, since those three values add up to less than 1. Someone who wants to identify signals with sets of worlds might identify s_i with $\{\omega_{i-1}, \omega_i, \omega_{i+1}\}$, as the set of worlds in which the signal is possible. But the optimal update rule disagrees with conditionalization on this proposition, or on any proposition in $\mathcal{F}$. If we imagine someone with extremely blurry vision, such that every signal is possible in every world, but more likely in some than others, this identification loses any appeal.

As Gallow notes, this result bears some similarity to Jeffrey conditionalization, in its modification by [13]. Every element of some partition (in this case, the partition into worlds ω) is scaled by some factor (in this case $L_\omega(s)$), and the resulting probability is renormalized.[26] Diaconis and Zabell [8] show that, even though Jeffrey's formalism was intended for circumstances in which no proposition was learned with certainty, every Jeffrey update is *as if* there were some proposition in a *larger* algebra that one learned with certainty. Our central result of this section shows that this is exactly how an agent should plan to update when there is some set of signals that they can respond to, that have probabilities in each world, even if the signals are not themselves expressible in terms of the σ-algebra the agent grasps over the worlds.[27]

This same set of results, including a connection between Jeffrey conditioning and updating by maximizing expected accuracy, is also derived by Konek [26]. He starts with a goal of fixing some issues for Jeffrey conditionalization. He does so by identifying S (in his notation, "$\mathcal{E}$"), as something that is only grasped in an experiential mode of presentation, though (like Carr) he thinks of it as grasped only *after* the learning experience has happened, rather than prospectively as something the agent can plan for. He explicitly thinks of L as a function describing the *objective chances* of having various experiences, while our framework might allow other interpretations.

An additional illustrative case is discussed by Isaacs and Russell (Section III). They discuss the case of someone who has arranged to meet an old friend at a cafe, whom they haven't seen in years. The person says, "When I see my friend, in fact I *know* it's my friend, and I have *evidence* that entails it's my friend. When I see a stranger, I don't know whether it's my friend or a stranger, and my evidence leaves both possibilities open." They say that in the bad case, when it isn't their friend, their "phenomenology is ambiguous", in a way that they sometimes recognize as not being their friend, but sometimes mistake for seeing their friend.

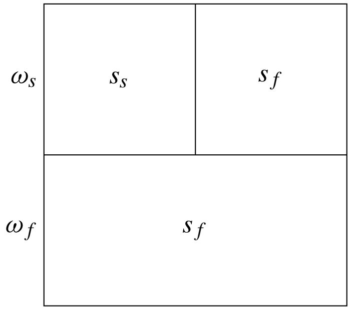
The way they end up modeling the case is as follows, in our notation. $\Omega = \{\omega_f, \omega_s\}$, representing the two possibilities that a friend or stranger might walk in, and p assigns each probability 1/2. $S = \{s_f, s_s\}$, representing the two responses they might have. The fact that they always get it right when their friend walks in is represented[28]

[26] For us, since this partition is the finest possible, the similarity to Jeffrey conditioning is purely formal, while for Gallow, the agent's credences over an algebra include propositions about themself and not just worldly propositions, and these really do shift in the Jeffrey way.

[27] Importantly, this means that our formalism for updating on one signal doesn't say anything about how the likelihoods of future signals might be affected. [15] raises a worry for [13], where repeating the same observation seems to "double-count" the results of that learning, which is not a result that we endorse. Saying how to properly deal with iterated updates is beyond the scope of this paper.

[28] It's worth noting that this model doesn't actually represent the *knowledge* the person gets in the friend case. But the important point that seems to be agreed by Williamson, and the others following him, is that the fact *that* this is actually knowledge is not usable by the agent. Since this model only aims to say how the agent should use whatever it is that they *can* use, in a way that maximizes expected accuracy, this model doesn't need to include the knowledge as a separate feature.

by $L_{\omega_f}(s_f) = 1$, and the ambiguity they get in the stranger case is represented by $L_{\omega_s}(s_s) = L_{\omega_s}(s_f) = 1/2$. As Isaacs and Russell show, the optimal update plan in this case has $P_{s_f}(\{\omega_f\}) = 2/3$, $P_{s_f}(\{\omega_s\}) = 1/3$, and $P_{s_s}(\{\omega_s\}) = 1$. It's precisely because $L_{\omega_f}(\{s_f\}) = 2 \cdot L_{\omega_s}(\{s_f\})$ that $P_{s_f}(\{\omega_f\}) = 2 \cdot P_{s_f}(\{\omega_s\})$.

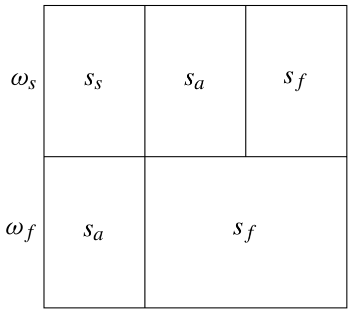


They describe the rule by saying, "When it's a stranger, be *certain* that it's a stranger; when it's your friend, have credence 2/3 that it's your friend." The idea is that they identify s_f with the person guessing that it's their friend, and therefore doing what they plan to do when it *is* their friend, and they identify s_s with the person guessing that it's a stranger, and therefore doing what they plan to do when it *is* a stranger. (This is similar to how Schultheis describes things as well.) But we don't have to interpret it this way. We can just interpret it as the person thinking that there's *some* binary distinction they'll be able to respond to, and trying to figure out how best to proportion their response to whatever it is that they're responding to. They might describe the distinction as "when it's your friend" versus "when it's a stranger", but this description is not an essential part of the situation.

To see that the description is not essential, we can consider a situation where a person has *three* possible signals they might experience when seeing the person: s_f and s_s as above, and also a more ambiguous s_a. Let us assume that $L_{\omega_f}(\{s_f\}) = 2/3$ and $L_{\omega_f}(\{s_a\}) = L_{\omega_s}(\{s_f\}) = L_{\omega_s}(\{s_a\}) = L_{\omega_s}(\{s_s\}) = 1/3$. That is, in the case where their actual friend walks in, they will probably receive the signal s_f, but might receive the signal s_a, but in the case where a stranger walks in, they could receive any of the three signals, all equally likely. In this case, because of the numbers we've given, P_{s_f} and P_{s_s} are the same as above, while $P_{s_a} = p$, because s_a is equally likely in each world. In this sort of case, we can imagine the person saying to themself "it's my friend!" or "it's a stranger." or "I can't quite tell.", but notably, these don't really correspond to worldly possibilities.[29] We don't commit ourselves to any claims about how these signals relate to knowledge that the agent has. Our main point is

[29] One might try to assign the first two to singletons of worlds, and the ambiguous signal to the pair of worlds, but we could tweak the case to add a fourth or fifth signal, and then this sort of move would be impossible, assuming there is no further proposition the agent grasps in a worldly way beyond the distinction between it being the friend and a stranger. Andrew Bacon argues (Ch. 6.3) that every such evidential role corresponds to a proposition, but for him this is a way of arguing that there are "vague propositions" that don't correspond to any set of worlds.

just that the update plan P that we describe is the one that maximizes expected accuracy.



Isaacs and Russell say (in Section V) that, on their picture, "there is *no* prior constraint on which doxastic plans you are allowed to have …but there is no guarantee that adopting such a plan will have the intended results." However, they do restrict plans to be a function from (guesses about) *worlds* to probability functions over worlds. We allow that plans might be a function from *signals* to probability functions over worlds, and there might well be more signals than worlds (as in our last example). They say that externalist considerations mean that "there may be no non-trivial rule that you can plan to follow with assurance of perfect success", and use this to explain why the trivial plan of always updating to the truth won't come out as having highest expected accuracy. But we think that this is insufficiently externalistic. There is *some* rule that the agent can plan to follow with assurance of perfect success — it just may not be determined by the propositions that the agent grasps in a worldly way, and may instead be determined by something that the agent grasps only as "whatever it is that I am able to respond to", which we are calling an "experiential" way of grasping. What is essential about the case we just described is that the agent has the capability of responding to their situation in three distinct ways, and that one of these responses is twice as likely as another (with the third impossible) in the case where they see their friend, and all three of these responses are equally likely in the case where they see a stranger. If this does not adequately describe the capabilities the agent has, then we should model the situation differently, with a set of signals that does adequately describe the capabilities the agent has.

The main result here is that as long as there is some set of entities, which we call "signals", that the agent is actually capable of responding to, and these entities have probabilities on each "world" that the agent distinguishes, then we have identified the unique update plan that maximizes expected accuracy out of all the plans that the agent is capable of following. We don't claim that the agent can conceptualize these signals in a worldly way. We don't claim that the agent has the meta-level capacity to represent the entire system as we do, to calculate the optimal plan. We don't claim, in Schultheis's terminology, that "adopting" the optimal plan is the same as following it. We do make the substantive assumption that every such function from signals to probability functions is a plan that the agent *can* follow. (This assumption is shared by Gallow and by Isaacs and Russell, given appropriate interpretations of "signals".)

And we claim that maximizing expected accuracy with respect to $p \times L$ makes a plan optimal.

5 Infinite Non-partitional Learning

Now we can put together the work of Sections 3 and 4 to show how to update by maximizing expected accuracy in the infinite non-partitional setting. As before, we have to apply measurability constraints to each function, and the philosophical motivation is the same — for the function to be epistemologically relevant, any information about the output that the agent is able to at least implicitly grasp must be determined by some sort of information about the input that the agent is able to at least implicitly grasp.

5.1 The Basic Result

The basic setting before we get to the signals is just as in the infinite partitional case, from Section 3. The rest of this paragraph is a review of assumptions from there. Ω is a set of worlds, and $\mathcal{F}$ is the σ-algebra of subsets of Ω representing the propositions the agent is able to implicitly grasp. This σ-algebra is countably-generated, because any physical being only directly grasps at most countably infinitely many propositions. The agent's prior $p\colon \mathcal{F} \to \mathbb{R}$ is a probability function, which requires that it is countably additive. $\Delta(\mathcal{F})$ is the set of all probability functions on this σ-algebra, and it comes equipped with its own σ-algebra $\mathcal{D}_{\mathcal{F}}$, which is generated by the sets $D_{A,x} = \{p \in \Delta(\mathcal{F}) : p(A) < x\}$, for any $A \in \mathcal{F}$ and $x \in \mathbb{R}$. $\mathcal{D}_{\mathcal{F}}$ is countably generated by the pairs where A ranges over members of some particular countable base for $\mathcal{F}$ and x ranges over the rational numbers. If $f\colon \Omega \to \mathbb{R}$ is measurable with respect to $\mathcal{F}$ over Ω and the Borel algebra $\mathcal{B}$ over $\mathbb{R}$, and has an upper bound, then $Exp_p(f)$ can be defined as the infimum of the expectations of the measurable functions greater than f that take on only finitely many distinct values, whose expectations are directly calculated by multiplying and adding. Accuracy is given by a function $U\colon \Delta(\mathcal{F}) \times \Omega \to \mathbb{R}$ that is bounded above, and measurable with respect to $\mathcal{D}_{\mathcal{F}}$, $\mathcal{F}$, and $\mathcal{B}$. We assume that U satisfies Strict Propriety, meaning that for any distinct $p, q \in \Delta(\mathcal{F})$, $Exp_p(U(p)) > Exp_p(U(q))$.

To deal with infinite non-partitional signals, we combine some of the work we did for infinite partitional signals with some of the work we did for finite non-partitional signals.

In the infinite partitional case, in Section 3.5, we had a set S of signals, and a function $L\colon \Omega \to S$. We defined a σ-algebra $\mathcal{S}$, the "informed σ-algebra", as the largest one that made L measurable. (We also worried that it might not be countably-generated, but assumed that it would be when a finite agent is able to grasp S.) An update plan was then a function $P\colon S \to \Delta(\mathcal{F})$ that was $(\mathcal{S}, \mathcal{D}_{\mathcal{F}})$-measurable. We defined $U(P)\colon \Omega \to \mathbb{R}$ by $U(P)(\omega) = U(P(L(\omega)), \omega)$, which is still measurable, as a composition of measurable functions. By assuming $\mathcal{S}$ was the largest possible, we made the measurability of P as weak a condition as possible. Restricting to a smaller algebra would in a sense be "leaving information on the table" in moving from Ω to S to $\Delta(\mathcal{F})$, and the update that maximized expected accuracy would typically have lower expected accuracy.

In the finite non-partitional case, in Section 4.3, we instead let $L\colon \Omega \to \Delta(\mathcal{S})$. We then used this to define $p \times L$ as a probability function over $\Omega \times S$ by multiplying the two together on singletons. When $P\colon S \to \Delta(\mathcal{F})$ was an update plan, we defined $U(P)\colon \Omega \times S \to \mathbb{R}$ by $U(P)(\omega, s) = U(P_s, \omega)$, and chose P to maximize $Exp_{p\times L}(U(P))$.

In the finite non-partitional case, we could assume that $\mathcal{S}$ consisted of all subsets of S, without comment, while in the infinite partitional case, we needed to construct $\mathcal{S}$ from $\mathcal{F}$ to make L measurable. But in the infinite non-partitional case, in order for L to be given at all, $\mathcal{S}$ must be given as well. It is just the class of propositions expressible in this "experiential" way, that the agent is somehow implicitly able to respond to, and that have likelihoods in various worlds. Since it is given by the nature of the agent and the situation, and not constructed in a way that might force it not to be countably-generated, we can safely assume it is in fact countably-generated. We will require that the update plan P is measurable with respect to this algebra, so that it doesn't depend on any experiential proposition the agent isn't able to respond to.[30]

There is a natural σ-algebra on $\Omega \times S$, which we will refer to as $\mathcal{F} \otimes \mathcal{S}$, generated by the sets $A \times T$, where $A \in \mathcal{F}$ and $T \in \mathcal{S}$. As long as $\mathcal{F}$ and $\mathcal{S}$ are countably-generated, so is this σ-algebra. Recall that $1_A\colon \Omega \to \mathbb{R}$ is defined by $1_A(\omega) = 1$ iff $\omega \in A$ and 0 otherwise. Note that, for fixed A and T, $1_A(\omega) \cdot L_\omega(T)$ is a bounded and measurable function that only depends on ω. Thus, since these $A \times T$ form a base for $\mathcal{F} \otimes \mathcal{S}$, we can define $p \times L\colon \mathcal{F} \otimes \mathcal{S} \to \mathbb{R}$ by letting $(p \times L)(A \times T) = Exp_p(1_A(\omega) \cdot L_\omega(T))$. The basic idea of this definition is the same as the one in the finite case, except that in the infinite case, it doesn't suffice to define the probability function on singletons. The idea is that the probability of A being true, while getting a signal of type T, is effectively the average probability of getting a signal of type T, over the worlds where A is true. (The definition we give is the standard way to extend a probability distribution from one space to a broader space using a "Markov kernel".)

As in the finite case, this probability function somehow mixes worldly and experiential propositions, so it is hard to interpret. But similarly, the formal condition for maximizing expected accuracy is now the one that we *would* have, if this probability function over the larger space $\Omega \times S$ had been the agent's prior, and they had been doing a partitional update on S. Thus, we can apply the formal results from Section 3 on infinite partitional learning, even though the functions involved are the combined ones rather than the worldly ones.

As before, when some signal s has positive probability (i.e., when $(p \times L)(\Omega \times \{s\}) > 0$), we get that $P_s(A) = \frac{(p\times L)(A\times\{s\})}{(p\times L)(\Omega\times\{s\})}$, for any update that maximizes expected

[30] This still makes the substantive assumption that the restrictions on possible update plans take the form of requiring features of the output to depend on features of the input, so that they can be expressed by measurability with respect to a σ-algebra. It is a question for further research whether other restrictions, like computability and complexity, yield interesting results about which plan is optimal.

accuracy. Additionally, Nielsen's result still applies, so P will maximize expected accuracy iff it is a regular conditional probability for $p \times L$ on $\mathcal{S}$.[31]

5.2 Using Densities to Calculate Updates

Knowing that the updates that maximize expected accuracy in the infinite, non-partitional setting are the regular conditional probabilities for $p \times L$ on $\mathcal{S}$ is not that helpful without a general means to present the prior p, the likelihood L, and a general method for calculating the regular conditional probability in terms of p and L. In the finite setting, p is easily presented in terms of its values on singletons, $p(\{\omega\})$, and L is easily presented in terms of its values on singletons, $L_\omega(\{s\})$. The update plan P that maximized expected accuracy was easily given in terms of its singletons,

$$P_s(\{\omega\}) = \frac{p(\{\omega\}) \cdot L_\omega(\{s\})}{\sum_{\omega' \in \Omega} p(\{\omega'\}) \cdot L_{\omega'}(\{s\})}.$$

For any given signal s, the denominator of the fraction is the same. This formula is structurally similar to Bayes' theorem. It basically says

> The update rule that sets the posterior of each world proportional to its prior, multiplied by the likelihood it gives the signal, maximizes expected accuracy.

In the infinite setting, probabilities on the singletons are not generally enough to specify full probability distributions, but in many useful examples, we can demonstrate a concept of "probability density" that effectively lets us do the same thing. Our Theorem 1 below effectively states:

> The update rule that sets the posterior density of each world proportional to its prior density, multiplied by the likelihood density it gives the signal, maximizes expected accuracy.

Readers who don't want to work through the formal definition of density, and the precise statement of the theorem, can skip directly to Section 5.3 to see the applications. Meehan & Zhang[28] derive this same rule as the result of applying some of Jeffrey's formal assumptions in infinite contexts, but we derive it as the way to maximize expected accuracy.

Let $\mathcal{X}$ be a countably-generated σ-algebra on X, and let $q : \mathcal{X} \to \mathbb{R}$ be a probability function. For any $A \in \mathcal{X}$, let $1_A : X \to \mathbb{R}$ be the indicator function, with $1_A(x) = 1$ iff $x \in A$ and 0 otherwise. Let $d : X \to \mathbb{R}$ be an $\mathcal{X}$-measurable function that is always non-negative, with $Exp_q(d(x)) = 1$. Then we can define $q' : \mathcal{X} \to \mathbb{R}$ by

$$q'(A) = Exp_q(1_A(x) \cdot d(x)),$$

[31] Strictly speaking, $(p \times L) : \mathcal{F} \otimes \mathcal{S} \to \mathbb{R}$ is a probability on $\Omega \times S$, so a regular conditional probability for it should be a family of probability functions on the same space, and the conditioning algebra should be an algebra on this whole space. However, $\mathcal{S}$ can be identified with the algebra $\{\Omega \times B : B \in \mathcal{S}\}$, which is a sub-$\sigma$-algebra of $\mathcal{F} \otimes \mathcal{S}$. Additionally, for any $A \in \mathcal{F}$, since $\mathcal{S}$ is countably generated, the regular conditional probability P_s will have $P_s(\Omega \times \{s\}) = 1$, and we can abuse notation and write $P_s(A)$ for $P_s(A \times \{s\})$, and thus think of P_s as a probability on Ω rather than on $\Omega \times S$.

and it is not hard to check that q' is a probability function. When q, q', d are related in this way, we say that d is a *density* for q' on q. Not every probability on $\mathcal{X}$ can be given in terms of a density on q[32] but many natural examples can, and we will use these to illustrate the theory.

In all of the cases we consider, p will be presented directly, usually as the uniform probability distribution on some subset of $\mathbb{R}^n$. Additionally, S will be isomorphic to some subset of $\mathbb{R}$, $\mathcal{S}$ will be the Borel σ-algebra on this set, and we will use q that is the uniform probability on this set. We will present the likelihood functions L_ω by giving a family of densities. In particular, we will define a non-negative measurable function $d\colon \Omega \times S \to \mathbb{R}$ with $Exp_q(d(\omega, s)) = 1$[33] for each ω, and the related families of functions $d_\omega\colon S \to \mathbb{R}$ given by $d_\omega(s) = d(\omega, s)$ and $d^s : \Omega \to \mathbb{R}$ given by

$$d^s(\omega) = \frac{d(\omega, s)}{Exp_p(d(\omega, s))},$$

or $d^s(\omega) = 1$ if $Exp_p(d(\omega, s)) = 0$. (We divide by the expectation in order that $Exp_p(d^s(\omega)) = 1$ for each s, so that d^s can be a density.) We will then apply the following theorem. Readers who are uninterested in the proof of this theorem can skip to Section 5.3 to see the applications.

Theorem 1 *With the above definitions, if d is bounded, and each d_ω is a density for L_ω on q, and each d^s is a density for P_s on p, then P is a regular conditional probability for $p \times L$ on $\mathcal{S}$.*

Proof The proof is essentially an exercise in several applications of Fubini's theorem, which states that if we have probability distributions over two variables, and a bounded measurable function of those two variables, then we get the same result if we compute the expectation of the function according to these two distributions in either order.

To state that P is a regular conditional probability for $p \times L$ on $\mathcal{S}$ means that

For every $A \in \mathcal{F}$ and $T \in \mathcal{S}$, $Exp_{p\times L}\left(1_T(s) \cdot P^A(s)\right) = (p \times L)(A \times T)$.

As mentioned in footnote 31, the expectation on the left is computed with respect to a probability distribution on $\Omega \times S$, while the function inside depends only on S. This is thus technically an abuse of notation, and we should formally think of these as functions of (ω, s) pairs that just don't depend on ω. However, we can also replace $(p \times L)$ by an equivalent probability distribution over just S. Define $q'\colon \mathcal{S} \to \mathbb{R}$ by $q'(T) = (p \times L)(\Omega \times T)$. So the left side is $Exp_{q'}(1_T(s) \cdot P^A(s))$. We use the following lemma (whose proof comes after the end of the proof of the theorem):

Lemma 1 *If d is bounded, and each d_ω is a density for L_ω on q, and $g\colon S \to \mathbb{R}$ is a bounded measurable funtion, then $Exp_{q'}(g(s)) = Exp_q(g(s) \cdot Exp_p(d(\omega, s)))$.*

[32] For instance, if d is a density for q' on q, and $q(A) = 0$, then $q'(A) = 0$. So if q' assigns a non-zero probability to any set that q assigns zero probability to, then q' can't be given in terms of a density on q. In fact, the Radon-Nikodym theorem states that this is the only reason such a density might not exist.

[33] Note that here we calculate the expectation of $d(\omega, s)$ with respect to q, which is a probability distribution only over s. We implicitly treat ω as a constant here and consider $d(\omega, \cdot)$ as a function from S to $\mathbb{R}$, and similarly whenever analogous expectations arise later on.

Returning to the statement that P is a regular conditional probability for $p \times L$ on $\mathcal{S}$, since 1_T and P^A are both bounded by 1, we can apply the lemma to see that the left side is

$$Exp_q(1_T(s) \cdot P^A(s) \cdot Exp_p(d(\omega, s))).$$

Recall that $P^A(s) = P_s(A)$. Since we assumed that each d^s is a density for P_s on p, we see that $P_s(A) = Exp_p(1_A(\omega) \cdot d^s(\omega))$. Recall that $d^s(\omega) = \frac{d(\omega,s)}{Exp_p(d(\omega,s))}$. Thus, $P_s(A) \cdot Exp_p(d(\omega, s)) = Exp_p(1_A(\omega) \cdot d(\omega, s))$. So the left side is

$$Exp_q(1_T(s) \cdot Exp_p(1_A(\omega) \cdot d(\omega, s)).$$

Now we calculate the right side. By definition, $(p \times L)(A \times T) = Exp_p(1_A(\omega) \cdot L_\omega(T))$. Because each d_ω is a density for L_ω on q, $L_\omega(T) = Exp_q(1_T(s) \cdot d(\omega, s))$. Thus, the right side is

$$Exp_p(1_A(\omega) \cdot Exp_q(1_T(s) \cdot d(\omega, s))).$$

Thus, both sides of the relevant equation are given by expectations of $1_A(\omega) \cdot 1_T(s) \cdot d(\omega, s)$, with respect to p and to q, calculated in different orders. All of the relevant functions are bounded and measurable, so by Fubini's theorem, these expectations are equal. □

Finally, we prove Lemma 1.

Proof We defined $q'(T) = (p \times L)(\Omega \times T)$ and want to show that

$$Exp_{q'}(g(s)) = Exp_q(g(s) \cdot Exp_p(d(\omega, s))),$$

whenever g is a bounded measurable function. We prove this equation holds whenever g is a function that takes on only finitely many distinct values, which then implies that it holds for all bounded measurable functions because of how expectations are defined.

So let g_i range over the finitely many values that g attains, and let $T_i = \{s \in S : g(s) = g_i\}$. Then $Exp_{q'}(g(s)) = \sum(g_i \cdot q'(T_i))$, and also $g(s) = \sum(g_i \cdot 1_{T_i}(s))$. By definition, $q'(T_i) = (p \times L)(\Omega \times T_i)$. By definition of $(p \times L)$, this yields $q'(T_i) = Exp_p(L_\omega(T_i))$. Since each d_ω is a density for L_ω on q, this yields $q'(T_i) = Exp_p(Exp_q(1_{T_i}(s) \cdot d(\omega, s)))$. By Fubini's theorem, since everything here is bounded, $q'(T_i) = Exp_q(1_{T_i}(s) \cdot Exp_p(d(\omega, s)))$.

Thus, $Exp_{q'}(g(s)) = \sum(g_i \cdot Exp_q(1_{T_i}(s) \cdot Exp_p(d(\omega, s))))$. Since scalar multiplication and addition distribute over expectations, we have $Exp_{q'}(g(s)) = Exp_q(\left(\sum g_i \cdot 1_{T_i}(s)\right) \cdot Exp_p(d(\omega, s)))$. Since $g(s) = \sum(g_i \cdot 1_{T_i}(s))$, we thus have

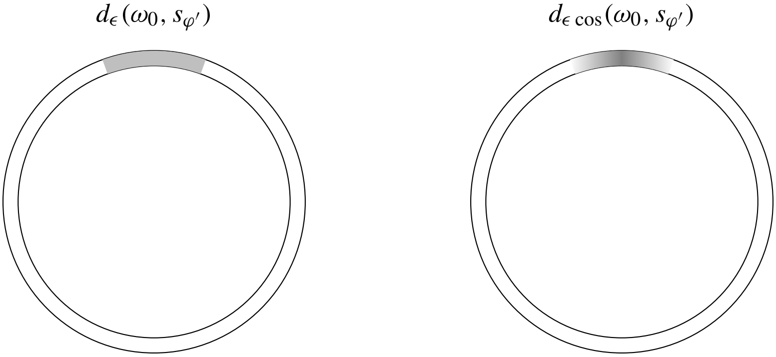
$$Exp_{q'}(g(s)) = Exp_q(g(s) \cdot Exp_p(d(\omega, s))),$$

as desired. □

5.3 Examples

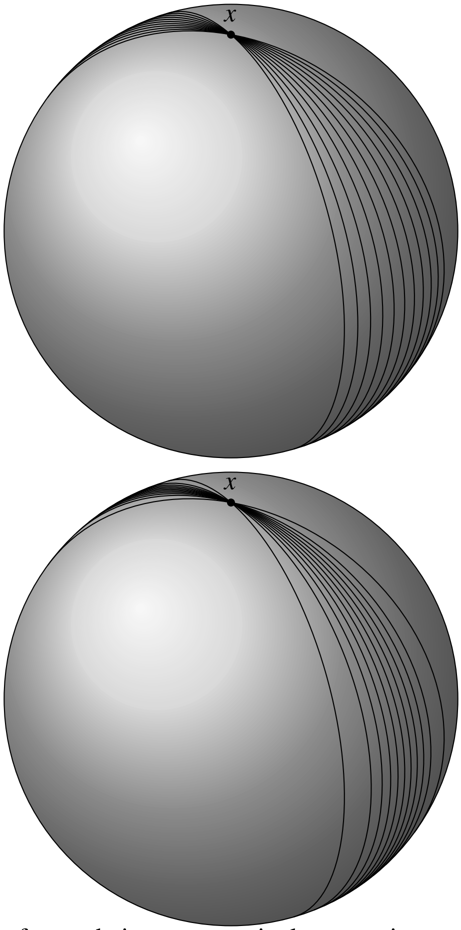
As a first example, we consider the case of a continuous unmarked dial, as opposed to the discrete unmarked clock we considered in Section 4.5. Let $\Omega = \{\omega_\varphi : 0 \leq \varphi < 2\pi\}$, and let p be any prior distribution over this dial. The agent is about to glance at the dial, and conceives of a continuum of ways that it might appear, which we represent as $S = \{s_\varphi : 0 \leq \varphi < \pi\}$.

For a first version, we might assume that the agent thinks of themself as having a precise ϵ margin for error, and uniform probability of updating on any signal within that interval from the true position of the dial. That is, they treat L_{ω_φ} as given by the density $d_\epsilon(\omega_\varphi, s_{\varphi'}) = \frac{1}{2\epsilon}$ iff $|\varphi - \varphi'| < \epsilon$ and 0 otherwise, where subtraction is considered modulo 2π. Because this density for each signal $s_{\varphi'}$ is the same at any world ω_φ with $|\varphi - \varphi'| < \epsilon$, the update is equivalent to conditionalizing on this set of worlds.



However, a more interesting version of the case uses a non-uniform probability of receiving signals within the margin for error. For instance, we might use $d_{\epsilon \cos}(\omega_\varphi, s_{\varphi'}) = \frac{\pi}{4\epsilon} \cos \frac{\pi}{2} \frac{\varphi - \varphi'}{\epsilon}$ iff $|\varphi - \varphi'| < \epsilon$ and 0 otherwise. This corresponds to a situation where the signal one receives is spread out around the actual world in a way that fades away to zero at distance ϵ, instead of being spread uniformly with a sharp cutoff. In this case, the update that maximizes expected accuracy shifts the prior in proportion to this density. On receiving signal s'_φ, the prior is altered in a way that raises probabilities of worlds ω_φ with φ near φ' but fades off to zero as $|\varphi - \varphi'|$ approaches ϵ. This update is not equivalent to conditionalization on any proposition in $\mathcal{F}$.

Another useful example returns to the Borel-Kolmogorov Paradox from Section 3.6. As in that example, we can let $\Omega = \{\omega_{\varphi,\theta} : 0 \leq \varphi < \pi, -\pi \leq \theta < \pi\}$, where φ represents longitude (in the sense involving full great circles) and θ represents latitude (in the sense where the south pole has "latitude" $-\pi$ and π, the north pole has "latitude" 0, the equator has "latitude" $\pm\pi/2$, and points in the western hemisphere have negative "latitude" while points in the eastern hemisphere have positive "latitude"). The agent starts with some prior p about where the point is located on the surface of the sphere. There is a measuring device that deterministically measures the longitude of the point, and displays it on an unmarked dial, which the agent will observe as in the previous example (but with $0 \leq \varphi < \pi$ rather than 2π). The signals $S = \{s_\varphi : 0 \leq \varphi < \pi\}$ represent the possible experiences the person could have on observing this dial. Because the measuring device is deterministic, the likelihood function over signals is given entirely by the agent's lack of precision in their observation. Thus, it is reasonable to use either of the forms d_ϵ or $d_{\epsilon\,\cos}$ from above, with the understanding that in this setting, the first input is $\omega_{\varphi,\theta}$ rather than ω_φ, but the likelihood densities still only depend on φ.



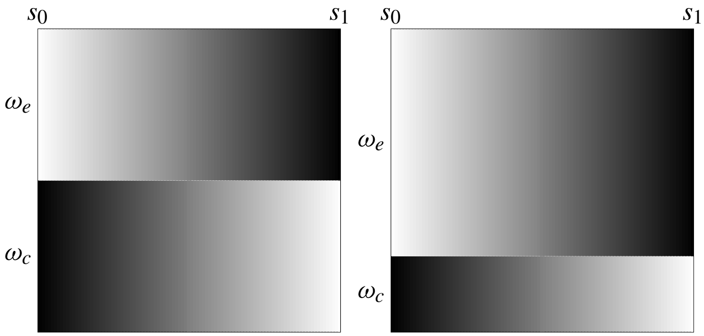
The posterior after updating on a particular s_φ assigns equal probability between these longitudes, for d_ϵ on the top, and $d_{\epsilon\,\cos}$ on the bottom.

The update that maximizes expected accuracy is a Jeffrey shift on the longitudes that acts proportionally to the likelihood of the experienced signal given the longitude, applied equally at every latitude. In Section 3.6, when the agent learned for sure what the precise longitude was, there was a somewhat unexpected factor applied to latitudes causing concentration near the equator — even though the prior distribution on the sphere was uniform, the posterior of having latitude between θ_1 and θ_0 was proportional to $(\cos\theta_0 - \cos\theta_1)/2$ rather than to $\theta_0 - \theta_1$. But in this case, where the agent remains uncertain what longitude the point is at, our result shows that the posterior probability density remains independent of latitude — the posterior *density* is shifted by a factor depending only on φ, but not θ. Concentration of the posterior itself on latitudes θ near the equator is inherited from the fact that any sector of the sphere spanning a given set of φ is wider near the equator. In the precise case, we needed this factor, because a single line of longitude doesn't get "thicker" near the equator. The latitude-dependence of the posterior in the Borel-Kolmogorov Paradox disappears when the learning is imprecise in this way — even though a given great circle is a longitude with respect to multiple different poles, a fixed "margin for error" of measurement (whether uniform or cosine) leads to uncertainty that is concentrated near the different equators. We think that learning situations involving infinitely many possible signals may always involve this kind of imprecision in realistic cases. If so, this defusing of the Borel-Kolmogorov paradox may be an important implication of our results in the infinite non-partitional setting.

One final example helps show that the set of signals can differ very significantly from anything the agent is able to grasp in a worldly way. Imagine that the agent is about to taste some coffee. They know that the coffee is either from Ethiopia or from Costa Rica, but they don't know which. They know that coffee from Ethiopia tends to have acidic, fruity notes, while coffee from Costa Rica tends to have more nutty notes. They suspect that many factors they don't conceive might play a role in determining how the coffee tastes when they sip it — those of us outside the problem might conceive of some of these as how the beans were roasted, how finely they were ground, temperature of the brewing water, the type of filter that was used, whether they have recently eaten something else that left traces in their mouth, and perhaps many other things. But the agent inside the problem has no idea how any of these factors might play a role, and thus doesn't consider them. They have tasted various other coffees and thus have a sense of what the nutty to fruity dimension is like.

We claim that it is natural to represent the situation in the following way. Ω just consists of two elements, $\{\omega_c, \omega_e\}$, since the only meaningful worldly distinction the agent can draw is between the coffee being from Costa Rica or from Ethiopia. However, the set S of signals may well be continuous, $\{s_x : 0 \leq x \leq 1\}$, where s_1 is the nuttiest taste experience while s_0 is the most acidic and fruity taste experience.

Let's assume the probability density of signal s_x is proportional to x conditional on ω_c, and proportional to $1 - x$ conditional on ω_e. If the agent starts with the prior $p(\omega_c) = p(\omega_e) = 1/2$, then the update plan that maximizes expected accuracy will have $P_{s_x}(\omega_c) = x$. That is, their posterior degree of confidence that the coffee is from Costa Rica is directly proportional to the nuttiness of their taste experience, as one might expect. If their prior had $p(\omega_c) = 1/4$ and $p(\omega_e) = 3/4$, then the update plan that maximizes expected accuracy will instead have $P_{s_x}(\omega_c) = \frac{x \cdot 1/4}{x \cdot 1/4 + (1-x) \cdot 3/4} = \frac{x}{3-2x}$. The optimal update is not identical to conditionalization on any worldly proposition (very clear since there are infinitely many possible signals but only finitely many worldly propositions), but it can be calculated simply from the prior, and the likelihood of the signals.



6 Conclusion

Greaves & Wallace [16] show, in the finite partitional setting, that the update plan that maximizes expected accuracy is the plan of conditionalization. We have provided a unified formalism that enables the relaxation of both the assumption of finiteness and the assumption of partitionality. In the infinite partitional setting, this formalism agrees with that of [10, 29]. But we think that we have provided more philosophical motivation for the treatment of σ-algebras that those papers use. In the finite non-partitional setting, this formalism generalizes those of [14, 19, 34]. We expect there to be some controversy about how we have treated the set S of "signals", that generalizes their use of "evidence" or "guesses". We claim that there is some set of signals that the agent is capable of responding to, even though the agent may well have no way of grasping propositions about these signals, other than by the experiential or practical mode, as "*that* type of signal". The fact that [1, 26] endorsed the same update plans, starting with a different set of assumptions, seems suggestive that there is something valuable here even for philosophers that disagree with our assumptions about signals. When we put these two generalizations together, we show that the update plan that maximizes expected accuracy in the infinite non-partitional setting is the same one that [28] derives from generalizing some formal properties of Jeffrey conditionalization. We show some methods for the theorist to calculate the update plan that maximizes expected accuracy, given the agent's prior and the likelihoods of various signals in various worlds. We show that in many contexts, this update avoids the Borel-Kolmogorov paradox, which arises in the infinite, partitional setting. We don't claim that the agent is capable of recognizing this plan as the one that maximizes expected accuracy, if they don't have this information about their own prior and these likelihoods. But if these signals are in fact the ones that the agent is capable of responding to, then this plan is in fact available to them, and does maximize expected accuracy out of that set.[34]

Author Contributions Both authors discussed the ideas of all sections of the manuscript, and were involved in the writing of all sections

[34] We would like to thank audiences at the 2023 Presidential Conference on Formal Epistemology, at the National University of Singapore, and the Rutgers Seminar in History and Foundations of Probability and Statistics, as well as Jeremy Goodman, Mikayla Kelley, Miriam Schoenfield, and Snow Zhang, for feedback on earlier versions of this article. We would also like to thank Bryan Mann for help with some of the images.

Funding Not applicable

Data Availability Statement Not applicable

Declarations

Ethical Approval Not applicable

Competing Interests None

References

1. Bacon, A. (2018). *Vagueness and Thought*. Oxford, UK: Oxford University Press.
2. Blackwell, D. (1951). Comparison of experiments. In Neyman, J., editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 93–102. University of California Press, Berkeley.
3. Blackwell, D. (1953). Equivalent comparisons of experiments. *The Annals of Mathematical Statistics, 24*(2), 265–272.
4. Blackwell, K., & Drucker, D. (2019). When propriety is improper. *Philosophical Studies, 176*(2), 367–386.
5. Briggs, R. (2009). Distorted reflection. *Philosophical Review, 118*(1), 59–85.
6. Carr, J. R. (2021). A modesty proposal. *Synthese, 198*(4), 3581–3601.
7. Das, N. (2022). Externalism and exploitability. *Philosophy and Phenomenological Research, 104*(1), 101–128.
8. Diaconis, P., & Zabell, S. L. (1982). Updating subjective probability. *Journal of the American Statistical Association, 77*(380), 822–830.
9. Dubra, J., & Echenique, F. (2004). Information is not about measurability. *Mathematical Social Sciences, 47*(2), 177–185.
10. Easwaran, K. (2013). Expected accuracy supports conditionalization–and conglomerability and reflection. *Philosophy of Science, 80*(1), 119–142.
11. Easwaran, K. (2013). Why countable additivity? *Thought: A Journal of Philosophy, 2*(1), 53–61.
12. Easwaran, K. (2019). Conditional probability. In R. Pettigrew & J. Weisberg (Eds.), *Open Handbook of Formal Epistemology, chapter 4* (pp. 131–198). PhilPapers.
13. Field, H. (1978). A note on jeffrey conditionalization. *Philosophy of Science, 45*(3), 361–367.
14. Gallow, J. D. (2021). Updating for externalists. *Noûs, 55*(3), 487–516.
15. Garber, D. (1980). Discussion: field and jeffrey conditionalization. *Philosophy of Science, 47*(1), 142–145.
16. Greaves, H., & Wallace, D. (2006). Justifying conditionalization: conditionalization maximizes expected epistemic utility. *Mind, 115*(459), 607–632.
17. Hervés-Beloso, C., & Monteiro, P. K. (2013). Information and σ-algebras. *Economic Theory, 54*(2), 405–418.
18. Hild, M. (1998). Auto-epistemology and updating. *Philosophical Studies, 92*(3), 321–361.
19. Isaacs, Y., & Russell, J. S. (2022). Updating without evidence. *Noûs*, page Forthcoming.
20. Jackson, F. (1982). Epiphenomenal qualia. *The Philosophical Quarterly, 32*(127), 127–136.
21. Jackson, F. (1986). What mary didn't know. *The Journal of Philosophy, 83*(5), 291–295.

22. Joyce, J. M. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science, 65*(4), 575–603.
23. Kadane, J. B., Schervish, M. J., & Seidenfeld, T. (1986). Statistical implications of finitely additive probability. In P. K. Goel & A. Zellner (Eds.), *Bayesian Inference and Decision and Techniques* (pp. 59–76). Amsterdam: Elsevier.
24. Kelley, M., & Neth, S. (2023). Accuracy and infinity: a dilemma for subjective bayesians. *Synthese, 201*(12), 1–14.
25. Kolmogorov, A. N. (1950). *Foundations of the Theory of Probability*. Chelsea Publishing Company.
26. Konek, J. (2023). The art of learning. In *Oxford Studies in Epistemology*, volume 7. Oxford University Press, Oxford, UK.
27. Lindley, D. V. (1982). Scoring rules and the inevitability of probability. *International Statistical Review, 50*(1), 1–11.
28. Meehan, A., & Zhang, S. (2020). Jeffrey meets kolmogorov: a general theory of conditioning. *Journal of Philosophical Logic, 49*(5), 941–979.
29. Nielsen, M. (2022). Regular conditional probabilities and strictly proper loss functions. *Statistics & Probability Letters, 185*, 109412.
30. Pettigrew, R. (2013). A new epistemic utility argument for the principal principle. *Episteme, 10*(1), 19–35.
31. Pruss, A. R. (2022). Accuracy, probabilism and bayesian update in infinite domains. *Synthese, 200*(6), 1–29.
32. Rescorla, M. (2015). Some epistemological ramifications of the Borel-Kolmogorov paradox. *Synthese, 192*, 735–767.
33. Schoenfield, M. (2017). Conditionalization does not (in general) maximize expected accuracy. *Mind, 126*(504), 1155–1187.
34. Schultheis, G. (2023). Accurate updating. Unpublished Manuscript.
35. Seidenfeld, T., Schervish, M. J., & Kadane, J. B. (2001). Improper regular conditional distributions. *The Annals of Probability, 29*(4), 1612–1624.
36. Stanley, J., & Williamson, T. (2001). Knowing how. *The Journal of Philosophy, 98*(8), 411–444.
37. Stinchcombe, M. (1990). Bayesian information topologies. *Journal of Mathematical Economics, 19*(3), 233–253.
38. Tóbiás, Á. (2021). A unified epistemological theory of information processing. *Theory and Decision, 90*, 63–83.
39. Williamson, T. (2011). Improbable knowing. In T. Dougherty (Ed.), *Evidentialism and Its Discontents* (pp. 147–164). Oxford, UK: Oxford University Press.