# Justifying the Norms
# of Inductive Inference

## Olav Benjamin Vassend

Bayesian inference is limited in scope because it cannot be applied in idealized contexts where none of the hypotheses under consideration is true and because it is committed to always using the likelihood as a measure of evidential favouring, even when that is inappropriate. The purpose of this article is to study inductive inference in a very general setting where finding the truth is not necessarily the goal and where the measure of evidential favouring is not necessarily the likelihood. I use an accuracy argument to argue for probabilism and I develop a new kind of argument to argue for two general updating rules, both of which are reasonable in different contexts. One of the updating rules has standard Bayesian updating, Bissiri *et al.*'s ([2016]) general Bayesian updating, Douven's ([2016]) IBE-based updating, and my (Vassend ([forthcoming])) quasi-Bayesian updating as special cases. The other updating rule is novel.

**ORCIDs:** Vassend, https://orcid.org/0000-0002-5964-8835.

## 1. Introduction

Bayesians hold that inductive inference requires two ingredients. First, a prior probability function defined on the hypotheses under consideration. Second, a likelihood function, which assigns a probability to the evidence conditional on each hypothesis. Intuitively, the prior probability assigned to a hypotheses represents how plausible it is that the hypothesis is true before the evidence has been taken into account. The likelihood, on the other hand, is a measure of evidential favouring: if $H_1$'s likelihood on the evidence is greater than $H_2$'s likelihood on the same evidence, then the evidence favours $H_1$ over $H_2$. Given a prior and likelihood, Bayesians hold that the prior probability of each hypothesis should be updated to a posterior probability through the use of Bayes's formula, so that the posterior probability of $H$ is proportional to the prior probability of $H$ multiplied by its likelihood.

Bayesianism has become the most common formal framework used by philosophers of science to study scientific methodology, and it is also an influential framework for statistical inference. But it rests on an assumption that is often violated in

scientific practice, namely, that one of the hypotheses under consideration is true.[1] Suppose none of the hypotheses under consideration is true, so that the goal is instead to find the hypothesis that is—in some sense—best. Depending on what is meant by 'best', the likelihood may not be an appropriate measure of evidential favouring. For example, suppose the goal is to identify the hypothesis whose expected maximal prediction error on future data is as low as possible. Then, as I show (Vassend ([forthcoming])), the likelihood is not an appropriate measure of evidential favouring because the hypothesis that has the best likelihood score on the evidence will in general not be the hypothesis that has the lowest expected maximal prediction error on future data. In this context, a more reasonable measure of evidential favouring may be one according to which the evidence favours $H_1$ over $H_2$ if and only if $H_1$'s maximal prediction error on the evidence is lower than $H_2$'s maximal prediction error on the evidence. The fact that Bayesianism is tied to using the likelihood as a measure of evidential favouring is therefore a limitation of the framework.

The goal of this article is to study inductive inference in a very general setting. Suppose our goal is to identify the best hypothesis $H$ (where 'best' does not necessarily mean 'true'). Let $p$ be a function that assigns a number between zero and one (inclusive) to each hypothesis, such that $p(H)$ is interpreted as representing a prior judgement of how plausible it is that $H$ is best (in the relevant sense) out of the hypotheses under consideration. In the rest of the article, I will refer to any such function as a 'credibility function'. Suppose, moreover, that $\mathrm{Ev}[E|H]$ is an evidential measure that is sensible given the purpose at hand. Then the questions to consider are as follows: (1) What norms should $p$ obey? (2) How should $p(H)$ and $\mathrm{Ev}[E|H]$ be combined in order to produce a posterior score $p_E(H)$ that represents how plausible it is that $H$ is best in light of $E$ and the prior information?

As we will see, one of the standard Bayesian arguments for probabilism generalizes, so that—given widely applicable conditions—$p$ and $p_E$ ought to be probability functions. The more interesting results concern updating. I will show that depending on what the goal is, the prior probability function and evidential measure should be combined in one of the following two ways in order to produce a posterior probability:

Inferential Updating: Given evidential measure Ev and prior probability function $p$, update $p$ to the posterior $p_E$ by way of the following formula:

$$p_E(H) = \frac{\mathrm{Ev}[E|H]p(H)}{\sum_i \mathrm{Ev}[E|H_i]p(H_i)}.$$

Predictive Updating: Given evidential measure Ev and prior probability function $p$, update $p$ to the posterior $p_E$ by way of the following procedure:

---

[1] This limitation is well known, but often ignored. For discussion of the problem, see, for example, (Box [1980]; Bernardo and Smith [1994]; Forster and Sober [1994]; Forster [1995]; Key, Pericchi and Smith [1999]; Shaffer [2001]; Sprenger [2009]; Gelman and Shalizi [2013]; Walker [2013]; Vassend [2019]; Sprenger [forthcoming]).

Step 1: For each $i$, calculate $q(H_i) = p(H_i) + \text{Ev}[E|H_i]$.
Step 2: Transform $q$ to $p_E$ as follows: for each $i$, $p_E(H_i) = 0$ or $p_E(H_i) = q(H_i) + d$, where $d$ is the unique number such that $d$ is minimal and, for all $i$, $p_E(H_i) \geq 0$ and $\Sigma_i p_E(H_i) = 1$.

The justification for the names of the two updating procedures will become clearer later. Inferential updating is clearly a generalization of Bayesian updating. Indeed Bayesian updating is just inferential updating with the likelihood used as the measure of evidential favouring.[2] What separates inferential updating from predictive updating is the former rule's commitment to regularity: inferential updating will never assign a probability of zero to any hypothesis, whereas predictive updating typically will. In Section 4, we'll see that a commitment to regularity is sometimes reasonable and sometimes not.

The plan for the rest of the article is as follows: In Section 2, I sketch an argument for why any credibility function ought to be probabilistic, regardless of whether the goal is truth or something else. Since the argument is a straightforward adaptation of Pettigrew's ([2016]) accuracy argument for probabilism, the section is brief. In Section 3, I give characterizations of inferential and predictive updating from a set of plausible assumptions. The strategy is to divide inductive updating into two steps: in the first step, the prior plausibility of a hypothesis is combined with the hypothesis's score on the evidence according to some measure of evidential favouring in order to produce a posterior score. In the second step, the posterior scores are normalized so that they are probabilistic. As we'll see, the requirement that the combination step and normalization step commute in certain desirable ways, together with a few other plausible assumptions, result in the conclusion that the combination step and normalization step must both be either multiplicative or additive. The characterizations of inferential and predictive updating are then just a few short steps away. I end the article with a discussion of inferential and predictive updating, including their relationship to each other and to other updating rules.

## 2. Why Credibility Functions Should Be Probabilistic

Before we can show that credibility functions ought to be probabilistic, we need to get clearer on what this claim amounts to. Let **H** be a set of hypotheses and suppose the goal is to identify the hypothesis in **H** that is best rather than true (where 'best' can mean anything we like). One complication that arises when 'true' is replaced by 'best' is that whereas there is only one true hypotheses, there may be several that are best.[3]

---

[2] Predictive updating, on the other hand, may remind the reader of the alternative to Jeffrey conditionalization derived by Leitgeb and Pettigrew ([2010]). The two rules do indeed share several features in common, although they are also importantly different. In fact, it is possible to derive a special case of predictive updating by using a proof strategy that resembles the one in (Leitgeb and Pettigrew [2010]). The Leitgeb–Pettigrew updating rule has been criticized by Levinstein ([2012]) for having unintuitive features. Predictive updating has some of the same features, but in fact I think these features are desirable in a context in which predictive accuracy is the goal. For more discussion, see (Vassend [unpublished]).

[3] I thank Reuben Stern for pointing this out to me.

For example, if 'best' means 'having a minimal maximum expected prediction error', then there may be several hypotheses that are tied for best. Note, however, that this is more a theoretical possibility than a practical one, since it is quite unlikely that multiple hypotheses would have (say) exactly the same predictive accuracy score, especially if the number of hypotheses is large. I will henceforth assume that at most one hypothesis out of the hypotheses under consideration is best. Note that if we make this assumption, then the hypotheses will also be mutually exclusive in the sense that in any subset of hypotheses at most one hypothesis can be best.

Another theoretical possibility is that none of the hypotheses under consideration is best. This can, for example, happen if the hypothesis space is infinite and does not contain a single best hypothesis, but rather an infinite sequence of hypotheses in ascending order of goodness.[4] To preclude this possibility, we must also assume that at least one of the hypotheses under consideration is best.

Provided we make the above assumptions, then there is nothing mathematically or philosophically that prevents us from treating **H** as a sample space. That is, **H** consists of hypotheses that are exhaustive in the sense that one of the hypotheses is best and mutually exclusive in the sense that at most one of the hypotheses is best in any collection of hypotheses. Note also that there is a natural $\sigma$-algebra on **H**. More precisely, union (or disjunction) and intersection (or conjunction) are defined in the normal way, the identity element for conjunction (that is, the top element of the algebra) is **H**, and the complement (negation) of any set $A$ formed through unions and intersections of subsets of **H** is defined in the following way: $\neg A := \mathbf{H} - A$. The main difference from the definition given in most philosophical treatments of Bayesianism is that the top element is now **H** rather than the tautology. This makes a big interpretive difference, but no difference to the mathematics.

Given the above set-up, we can now define what it means for a function on the algebra, **H***, generated by **H** to be probabilistic in the following way:

> Probability Axioms: A function $p$ defined on **H*** is probabilistic if and only if it satisfies the following requirements:
>
>   (i)  $p(\mathbf{H}) = 1$.
>   (ii) $p(A) \geq 0$ for all subsets $A$ of **H***.
>   (iii) $p(A \lor B) = p(A) + p(B) - p(A\&B)$, for all subsets $A$ and $B$ of **H***.

Note that credibility functions automatically satisfy (ii) since we have defined them to have a range between zero and one, so the real question is whether they ought to satisfy (i) and (iii). One of the standard arguments for why regular credence functions (or degrees of belief) ought to be probabilistic is the accuracy argument (Joyce [1998], [2009]; Predd *et al.* [2009]; Pettigrew [2016]). Briefly, the argument is as follows[5]:

---

[4] I thank a referee for pointing out this possibility.
[5] There are several versions of the argument; here, I present a variant of Pettigrew's ([2016]) version.

the ideal credence function to have is the function that assigns one to the hypothesis that is true and zero to all hypotheses that are false. Suppose now that we have a divergence measure (satisfying certain reasonable properties) that quantifies the distance between the ideal function and any other candidate credence function. It can then be shown that any credence function that is not probabilistic will be dominated by some probabilistic function in the sense that the probabilistic function will be guaranteed to have a smaller divergence from the ideal function. Since it is irrational to choose an option that is known to be dominated, it follows that it is irrational to use a non-probabilistic credence function.

An interesting fact about the accuracy argument for probabilism is that it does not depend for its validity on any specific interpretation of the credence function, nor does it depend on the assumption that the ideal credibility function is the function that assigns one to the hypothesis that is true and zero to all hypotheses that are false. Indeed, nothing in the accuracy argument prevents us from designating the ideal credibility function otherwise. Hence, we can easily adapt the argument to a context where the goal is to identify the hypothesis that is best rather than true. In such a context, the ideal function would clearly be one that assigns one to the hypothesis that is best and zero to all other hypotheses. We can then formulate the following version of the accuracy argument:

(P1) The ideal credibility function is the function that assigns one to the hypothesis that is best and zero to all other hypotheses.

(P2) Given any non-probabilistic function, there is a probabilistic function that is guaranteed to have a smaller divergence from the ideal function (given that the divergence measure has certain reasonable properties).

(P3) Given any probabilistic function, there does not exist any function that is guaranteed to have a smaller divergence from the ideal function (given that the divergence measure has certain reasonable properties).

(P4) If P1–P3, then non-probabilistic credibility functions are irrational.

(C) Non-probabilistic credibility functions are irrational.

P2 and P3 are mathematical theorems (proven by Predd *et al.* [2009]) that hold regardless of what we choose as the ideal function. P1 and P4, on the other hand, are intuitively reasonable general rational principles. The main question that may be raised about the generalized version of the accuracy argument is whether the conditions on the divergence measure are still reasonable when truth is no longer the goal. For example, P2 and P3 require the assumption that the divergence measure belong to the class of Bregman divergences. Is this a reasonable requirement to make? My only response to this question is that I do not see how this assumption (and other necessary mathematical assumptions) are more plausible if truth is the goal than if the goal is to identify the hypothesis that is best in some other sense. So, at least in my eyes, the generalized accuracy argument is at least as plausible as the original argument. In any case, my main goal in this article is not to give a careful analysis of the accuracy argument. From now on I will assume that any credibility function ought to be probabilistic.

That is, I will assume that if $p$ is a function that assigns a number between zero and one to each hypothesis $H$ that represents how plausible it is that $H$ is best (in some sense), then $p$ ought to be probabilistic. In the next section, I turn to the main question of the article: given a probability function $p$ and given a piece of evidence $E$, how should $p$ be updated in light of $E$?

## 3. Deriving the Updating Rules

Suppose we have a credibility function defined on a hypothesis set **H** that is probabilistic in the sense of the preceding section. Suppose, also, that we have an evidential measure function $Ev[E|H]$ defined on the set of evidence and the set of hypotheses under consideration. Note that we are not assuming that $Ev[E|H]$ is probabilistic (for example, $\Sigma_i Ev[E|H_i]$ need not sum to one). It is widely accepted that if the goal is to find the true hypothesis in a partition of hypotheses and the evidential measure is the likelihood, $Ev[E|H] = p(E|H)$), then any probability function over the hypotheses ought to be updated through Bayesian updating:

Bayesian Updating: $p_E(H) = \dfrac{p(E|H)p(H)}{\Sigma_i p(E|H_i)p(H_i)}$.

The natural generalization of Bayesian updating is what I have called inferential updating in the introduction. However, it is not clear why the prior probability function and the evidential measure should always be combined in a Bayesian-like manner, regardless of what the evidential measure is and regardless of what the purpose of updating is. Unfortunately, whereas the accuracy argument for probabilism does not make any assumptions about how the credibility function is interpreted, the standard accuracy argument for Bayesian updating (Greaves and Wallace [2006]) relies on properties that are unique to the likelihood, in particular the fact that the likelihood forms a joint distribution with the prior. Thus, the standard accuracy argument does not generalize to cases where the evidential measure is not the likelihood. Other standard arguments for Bayesian updating have the same limitation (for example, Dutch book arguments). A different kind of approach is therefore needed.

Bissiri *et al.* ([2016]) come up with a different approach. They show that provided that the evidential measure is a function of an additive loss function, $L(E, H)$, such that $Ev[E_1 \ \& \ E_2|H] = f(L(E_1, H) + L(E_2, H))$, and given that a few other assumptions are met, then the updating procedure must have the following form, where $c$ is some constant:

$$p_E(H) = \frac{e^{-c*L(E|H)}p(H)}{\sum_i e^{-c*L(E|H_i)}p(H_i)}. \tag{1}$$

Bissiri *et al.* ([2016]) call the above updating procedure 'general Bayesian updating'. General Bayesian updating traces back to (Zhang [2006]) and has been increasingly

influential in statistics in recent years.[6] Although Bissiri *et al.*'s ([2016]) argument for general Bayesian updating is interesting, it has several limitations. One problem is that, as argued in (Vassend [2019]), the probabilities in Equation (1) cannot be interpreted in the standard Bayesian way as plausibilities of truth. But if the probabilities are not standard credibility functions, then the decision theoretic framework assumed by Bissiri *et al.* ([2016]) would seem to lack justification. The argument also makes certain mathematical assumptions that seem hard to justify from a philosophical point of view. In particular, the authors base their argument in part on the use of statistical divergence measures, and they assume that the divergence belongs to the class of $f$-diverences.[7] This assumption rules out many standard divergence measures, including all Bregman divergences aside from the Kullback–Leibler divergence (Amari [2009]).[8] A final limitation of Bissiri *et al.*'s ([2016]) derivation is that there are many reasonable evidential measures that cannot be written as a function of an additive loss function. Indeed, even the likelihood will only have such a form if the evidence is independent conditional on $H_i$, for all $i$.[9] Thus, although their argument is interesting, a more general approach that makes less restrictive and more philosophically defensible assumptions is desirable. That is the goal of this section. Later we will see that Bissiri *et al.*'s ([2016]) updating rule may be derived as a special case.

To start, note that ordinary Bayesian updating can be decomposed into two steps:

Combination Step: For each $i$, calculate $p^*(H_i) = p(E|H_i)p(H_i)$.

Normalization Step: Transform $p^*$ to $p'$ as follows: for each $i$, $p'(H_i) = \frac{p^*(H_i)}{p(E)}$.

In the first step, the prior plausibility of the hypothesis is combined with the evidential score (that is, the likelihood) of the hypothesis in order to produce an overall judgement of the hypothesis's posterior plausibility. In the second step, the posterior plausibility of all the hypotheses are rescaled in such a way that they jointly obey the probability axioms, such that all the posterior plausibility scores fall between zero and one, inclusive, and jointly sum to one.

Bayesian updating is a special case of a much broader class of updating rules that decompose into a combination step and a normalization step. The purpose of the remainder of this article will be to study this class of updating rules. The combination

---

[6] See (Grünwald and van Ommen [2017]) for a thorough discussion of general Bayesian updating and related updating rules.

[7] They also give an alternative derivation that does not make this assumption. However, the alternative derivation makes other suspect assumptions. In particular, it assumes that the normalization procedure is multiplicative, which we'll see later in this article can be put into question.

[8] Recall that Bregman divergences play a crucial role in the accuracy argument for probabilism. The justification for the focus on Bregman divergences is their tight connection to strict propriety (see Predd *et al.* [2009]).

[9] If $p(E_1, E_2|H) = p(E_1|H)p(E_2|H)$, we can write $p(E_1, E_2|H) = e^{\log p(E_1|H) + \log p(E_2|H)}$, that is the likelihood is of the form required by Bissiri *et al.* ([2016]). But if $p(E_1, E_2|H) \neq p(E_1|H)p(E_2|H)$, then we cannot write the likelihood in this way.

step requires a combination function, $c$, that takes as its input a prior probability, $p(H)$ and a set of evidential scores, $\text{Ev}[E_1|H]$, $\text{Ev}[E_2|H, E_1]$, $\text{Ev}[E_3|H, E_1, E_2]$, and so on, and that assigns a total score to $H$, taking into consideration both its prior probability and its performance on the evidence. The normalization step then transforms those scores into probabilities. In other words, on an abstract level, our purpose will be to study updating procedures that decompose in the following way:

> Combination Step: For each hypothesis, $H_i$, a set of evidential scores and a prior probability are combined using some combination function $c$ in order to produce an overall posterior score for $H_i$.

> Normalization Step: The posterior scores of all the $H_i$ are transformed using some function $N$ such that they jointly satisfy the probability axioms.

In the next two subsections the combination step and the normalization step are analysed in detail. The goal is to show that—given reasonable assumptions—the combination function $c$ and the normalization function $N$ both have a very limited set of possible functional forms.

### 3.1. The combination step

Let $e_1$ and $e_2$ represent the evidential scores of a hypothesis $H$ on some evidence, and let $h$ represent $H$'s prior probability; then there are two candidate forms for the combination function that arguably stand out as being particularly plausible:

> Additive Combination: $c(e_1, e_2, h) = e_1 + e_2 + h$.

> Multiplicative Combination: $c(e_1, e_2, h) = e_1 * e_2 * h$.

Note that $e_1$ and $e_2$ here may represent either conditional or unconditional evidential scores. For example, $e_1$ may represent $\text{Ev}[E_1|H]$, the unconditional evidential score of $H$ on $E_1$, or it may represent $\text{Ev}[E_1|H, E_2]$, the conditional evidential score of $H$ on $E_1$ given that $E_2$ has already been taken into account. Note, also, that to say that the combination function is additive or multiplicative is not the same as saying that the evidential measure is additive or multiplicative in the sense that $\text{Ev}[E_1, E_2|H] = \text{Ev}[E_1|H] + \text{Ev}[E_2|H]$ or $\text{Ev}[E_1, E_2|H] = \text{Ev}[E_1|H] * \text{Ev}[E_2|H]$. The latter assumptions are much stronger, and amount to assuming that $E_1$ and $E_2$ are independent conditional on $H$ (relative to the evidential measure Ev).

If we make a few reasonable assumptions, we can prove that the combination function must be multiplicative or additive. First of all, suppose we have evidential scores $e_1$ and $e_2$, and a prior probability $h$. Clearly, the order in which we combine the evidential scores and the prior should not matter for the final result we get. That is not to say that the order in which the evidence is received does not matter; it may. For example, if we flip a coin and the outcomes are six heads in a row and then six tails in a row, then the order of the outcomes strongly suggest that the outcomes are probabilistically dependent. Nevertheless, the order in which we evaluate the available pieces

of evidence in order to produce an overall judgement should not influence the overall judgement at which we arrive. For that reason, the combination function should be commutative: $c(e_1, e_2) = c(e_2, e_1)$. Furthermore, it clearly should not matter whether we first combine $e_1$ and $e_2$ and then combine the result of that with $e_3$, or whether we combine $e_2$ with $e_3$ and then combine the result with $e_1$, or whether we combine all three pieces of evidence at the same time. In other words, $c$ should be associative: $c(e_1, c(e_2, e_3)) = c(c(e_1, e_2), e_3) = c(e_1, e_2, e_3)$.

The final reasonable requirement is more quantitative. Clearly, the impact that $e_1$ has on $H$'s overall evidential score, after $e_2$ has already been taken into account, should not depend on the impact that $e_2$ has on $H$. That is not to say that a piece of evidence $E_2$ should not influence the impact that a different piece of evidence $E_1$ has on $H$'s evidential score; it may well, but if it does it should do so through $Ev[E_1 | H, E_2]$. A piece of evidence may influence the evidential impact conferred by another piece of evidence, but the evidential scores themselves should not influence each other. In other words, the requirement is that the impact that, for example, $e_1 = Ev[E_1 | H, E_2]$ makes on $H$'s total evidential score should not depend on the impact that $e_2 = Ev[E_2 | H]$ makes on $H$'s total evidential score, nor vice versa.

Given that we are willing to suppose that the combination function is twice differentiable, the preceding requirement may be naturally formalized as constraints on the partial derivatives of the combination function. Let $c(x, y)$ be the combination function as a function of variables $x$ and $y$. Then the impact that the evidential score $e_1$ makes on $H$'s total evidential score is plausibly the value of the partial derivative of $c(x, y)$ with respect to $x$, when evaluated at $x = e_1$. If $\frac{\partial c(x,y)}{\partial x} c(x = e_1, y)$ is a large number, then that means setting $x$ to $e_1$ makes a large difference to $H$'s overall evidential score; if it is zero, then $e_1$ makes no difference.

The requirement that the impact that $e_1$ makes should not depend on the impact that $e_2$ makes, nor vice versa, for any $e_1$ and $e_2$, may then be formalized in terms of a constraint on the higher-order partial derivatives of $c$, namely, that for some constant $k$ the following equation be obeyed:

$$\frac{\partial^2 c(x, y)}{\partial x \partial y} = k.$$

The above equation formalizes the idea that the impact that $x$ makes, $\frac{\partial c}{\partial x}$, should not depend on the impact that $y$ makes, $\frac{\partial c}{\partial y}$, where $x$ and $y$ represent any possible evidential scores. We can now show the following (the derivation is in the appendix):

> Characterization of the combination function: Suppose the combination function, $c(x, y)$ satisfies the following requirements:
>
> (1) $c$ is commutative.
> (2) $c$ is associative.
> (3) $c$ is twice differentiable.
> (4) $c$'s partial derivatives satisfy the following equation, for some number $k$:

$$\frac{\partial^2 c(x,y)}{\partial x \partial y} = k.$$

Then $c$ must have one of the following two forms:

(1) If $k = 0$, then $c(x,y) = x + y$.
(2) If $k \neq 0$, then $c(x,y) = xy$.

Hence, it follows that the combination function must be additive or multiplicative. Of course, this conclusion is only as plausible as the assumptions from which it is derived, and some people may be uncomfortable with some of the assumptions that have been made, in particular the condition on the partial derivatives of the combination function. As it happens, it's possible to derive the conclusion from quite different assumptions. Hence, in order to show the robustness of the conclusion, I provide an alternative characterization of the combination function in the appendix.

## 3.2. The normalization step

After the combination function has produced a posterior plausibility score, the posterior score must be normalized to be a probability. In theory, normalizing a set of numbers means transforming the numbers in such a way that they are all between zero and one and jointly sum to one, while at the same time retaining as much of their internal structure as possible. In practice, this means that the most extreme numbers in the set may be forced to take the value zero, while the remaining numbers in the set are rescaled by some function, $f$. In other words, normalization in general takes the following functional form:

$$N(x) = \begin{cases} 0 & \text{given that } x \text{ is sufficiently low,} \\ f(x) & \text{otherwise.} \end{cases} \tag{2}$$

For example, in the normalization step of standard Bayesian updating, $N(x) = f(x)$ (that is, no non-zero numbers are normalized to zero) and if the set to be normalized is $\{a_1, a_2, \ldots, a_n\}$, then $f(x) = \frac{1}{\sum_i a_i}$. Note that both $N$ and $f$ are relative to the set that is being normalized; hence, if we need to be precise, we should write $N_S$ and $f_S$, where the subscript indicates the set that is being normalized. Nevertheless, I will typically leave off the subscripts in order to avoid clutter.

Clearly, $f$ should be a one-to-one function. Indeed, except in the case where $x$ and $y$ are both normalized to zero, it should be the case that if $x < y$ then $f(x) < f(y)$. Furthermore, it is clear that the function $f$ ought to commute with the combination function. Suppose we have scores $e_1$, $e_2$, and $h$. Then we should arrive at the same posterior probability regardless of whether we do either of the following: first we combine $h$ and $e_2$, normalize, then combine the normalized result with $e_1$ and normalize again; or we first combine $h$ and $e_1$, normalize, and then combine that normalized result with $e_2$ before normalizing again. In symbols, we require, for all possible scores $x$, $y$, and $z$,

that: $f(c(x, f(c(y, z)))) = f(f(c(x, y), z))$. The justification for this requirement is, again, that the order in which we evaluate our evidence—which is arbitrary—should not have an influence on our final judgement. By combining just the preceding two requirements, we can show the following:

> Characterization of the Normalization Procedure: Suppose we have a normalization procedure as in Equation (2) that satisfies the following requirements:
>
> (1) $f$ commutes with the combination function $c$. For all $x$, $y$, and $x$: $f(c(x, f(c(y, z)))) = f(f(c(x, y), z))$.
> (2) $f$ is one-to-one: for all $x$ and $y$, $f(x) = f(y)$ if and only if $x = y$.
>
> Then the normalization process must have one of the following forms, for some constant $k$ that depends on the set, $S$, of numbers being normalized:
>
> (1) If the combination function is multiplicative, then, for all $x$ in $S$, $f(x) = k * x$.
> (2) If the combination function is additive, then, for all $x$ in $S$, $f(x) = x + k$.

The proof, which again is straightforward, is in the appendix.

## 3.3. Characterizations of inferential and predictive updating

The results so far show that any updating procedure needs to have either: (1) A multiplicative combination step and a multiplicative normalization step, or (2) an additive combination step and an additive normalization step. Call an updating procedure that satisfies either (1) or (2) a legitimate updating procedure.[10]

To characterize inferential updating we now introduce the following principle:

> Regularity: No hypothesis is ever conclusively ruled out by any evidence unless the evidence logically refutes the hypothesis; that is, the posterior probability of any hypothesis is always greater than zero.

We can then show the following (see the appendix):

> Characterization of Inferential Updating: The only legitimate updating procedure that satisfies regularity is inferential updating. That is, given evidential measure Ev and prior probability function $p$, update $p$ to the posterior $p_E$ by way of the following formula:

$$p_E(H) = \frac{\text{Ev}[E|H]p(H)}{\sum_i \text{Ev}[E|H_i]p(H_i)}.$$

---

[10] Note that not every updating rule that has been suggested in the literature is legitimate in this sense of the word. For example, Douven and Wenmackers ([2017]) consider a rule according to which $p_E(H) = c * (p(H) * p(E|H) + f(E, H))$ where $c$ is a normalization constant and $f(E, H)$ is a 'bonus' assigned to $H$ in case $H$ is the best explanation of $E$. This updating rule is not legitimate because it is neither purely additive nor purely multiplicative. On the other hand, the class of rules considered in (Douven [2016]) are legitimate.

Inferential updating satisfies regularity; it will never result in any hypothesis having a posterior probability of zero. On the other hand, in the appendix, I show that an updating procedure that uses an additive combination function and an additive normalization function must violate regularity; most of the time, an additive updating rule will assign a posterior probability of zero to some hypotheses. But this does not mean that such an updating rule should never be used. As we will see in the next section, sometimes we may want to be able to exclude certain hypotheses from consideration—that is, assign them a posterior probability of zero.

Nevertheless, we do not want to exclude more hypotheses than is warranted by the data. The updating procedure ought to be conservative and exclude as few hypotheses as possible at every step. In other words, any updating procedure that violates regularity should plausibly still satisfy the following principle:

> Conservativeness: The updating procedure assigns a posterior probability of zero to as few hypotheses as possible, given the combination function, the normalization procedure, and the evidence available.

We are now in a position to characterize predictive updating:

> Characterization of Predictive Updating: The only legitimate updating procedure that violates regularity, but satisfies conservativeness, is predictive updating. That is, given evidential measure Ev and prior probability function $p$, update $p$ to the posterior $p_E$ by way of the following procedure:
>
> Step 1: For each $i$, calculate $q(H_i) = p(H_i) + \text{Ev}[E|H_i]$.
>
> Step 2: Transform $q$ to $p_E$ as follows: for each $i$, $p_E(H_i) = 0$ or $p_E(H_i) = q(H_i) + d$, where $d$ is the unique number such that $d$ is minimal and, for all $i$, $p_E(H_i) \geq 0$ and $\Sigma_i p_E(H_i) = 1$.

## 4. Discussion of Inferential and Predictive Updating

### 4.1. The difference between inferential updating and predictive updating

Inferential updating and predictive updating differ in that the former updating rule obeys regularity while the latter rule does not. Is regularity a reasonable constraint? In some contexts it is, but in others it is not. Suppose our main priority is to identify the hypothesis that is true or (if none of the hypotheses is true) the hypothesis that is closest to the truth according to some appropriate measure of closeness to the truth. Given this goal, it is reasonable to be risk-averse and open-minded: we do not want to rule out any hypothesis as potentially being the hypothesis that is true. Even if a lot of evidence strongly suggests that a hypothesis is false, there is always the possibility that the evidence is unrepresentative or misleading. And so regularity is a reasonable constraint in this context.

However, suppose we do not care about which of our hypotheses is true or closest to the truth; our goal is not inferential, but predictive. We wish to find, as efficiently as

possible, the subset of hypotheses that can be expected to be as predictively accurate as possible. In this context, there is no theoretical justification for requiring that the updating rule obey regularity; on the contrary, there are good reasons for why we might want an updating rule that violates regularity. In particular, suppose the posterior distribution will be used in order to make a weighted probabilistic prediction, so that the goal is for $p(D|H_i)p_E(H_i)$ to be as accurate on future data $D$ as possible. In that case, it would seem inadvisable to assign positive probability to any hypothesis that has shown itself to be very predictively inaccurate, since the predictions made by such a hypothesis would likely throw off the weighted prediction. On the other hand, we do not want to go to the opposite extreme and base the prediction on the single hypothesis that has performed best on the evidence, as that is liable to lead to overfitting (Forster and Sober [1994]). Predictive updating enables one to set the probabilities of predictively inaccurate hypotheses to zero in a principled (and conservative) way.

Let's consider a specific example. When the hypotheses under considerations make probabilistic predictions and the goal is maximal predictive accuracy, it is natural to use a strictly proper scoring rule as the measure of evidential favouring (Gneiting and Raftery [2007]). For various reasons, the most popular scoring rule in applied research is probably the continuous ranked probability score (CRPS). Suppose we have a set of competing statistical models $M_1$, $M_2$, and so on, and for each model, let $p_{M_i}$ be the marginal (cumulative) probability forecast distribution corresponding to $M_i$. Suppose, moreover, that $p_{M_i}$ has finite first moment, that $X$, $X_1$ and $X_2$ are independent and identically distributed random variables that follow the distribution of $p_{M_i}$, and that $x$ is the actual observed outcome. Then the CRPS can be written in the following way (where the expectations are taken relative to $p_{M_i}$):

$$\text{CRPS}(p_{M_i}, x) \;=\; \text{E}|X - x| - \frac{1}{2}\text{E}|X_1 - X_2|. \tag{3}$$

As Equation (3) makes clear, CRPS is a statistical generalization of absolute error. As Gneiting and Raftery ([2007]) point out, a significant benefit of the CRPS is that it is easily interpretable, since the outputs of Equation (3) can be reported in the same units as the measurements. For example, suppose the measurements are in terms of meters. Then the CRPS score of a model on an observation will be a representation of how many meters inaccurate the model's predictions are of that observation, on average (since the prediction is a probability distribution rather than a single number, the average is needed).

If we let $\text{Ev}[x|p_{M_i}] = a * \text{CRPS}(p_{M_i}, x)$, where $a$ is some constant, and assign prior probabilities to all the models, then predictive updating can be used to assign posterior probabilities to all the models.[11] Importantly, given sufficient evidence (and depending

[11] If the models contain parameters, then the probability distributions over those parameters may be updated using either inferential or predictive updating.

on how the constant $a$ is chosen) many of the models will receive a posterior probability of zero. These posterior probabilities can then be used for model selection or for making a weighted prediction using all the models. Of course, it is an empirical question whether predictive updating is better (for predictive purposes) than inferential updating (including standard Bayesian updating). An empirical evaluating of predictive updating will have to wait for a different occasion, however. In this section I have simply tried to suggest one concrete way in which predictive updating may be implemented.

## 4.2. The relationship between inferential updating and other updating procedures

As was already mentioned in the introduction to the article, standard Bayesian updating is clearly a special case of inferential updating: more precisely, we get Bayesian updating if and only if $\text{Ev}[E|H] \propto p(E|H)$; that is, if and only if the evidential measure is proportional to the likelihood. What (Vassend ([forthcoming]) calls 'quasi-Bayesian updating' is also a special case of inferential updating; indeed, quasi-Baysian updating is simply inferential updating with an evidential measure that has been suitably calibrated to a verisimilitude measure. Similarly, Douven's ([2016]) IBE-based updating rule is also clearly a kind of inferential updating.

Perhaps more interestingly, Bissiri *et al.*'s ([2016]) general Bayesian updating is also a special case of inferential updating. More precisely, we have:

General Bayesian Updating Is a Special Case of Inferential Updating: Suppose the evidential measure Ev is a strictly decreasing function $f$ of some loss function, $L(E, H)$, such that for all $E_1$ and $E_2$, Ev satisfies the following conditions:

(1) $\text{Ev}[E_1|H, E_2] = \text{Ev}[E_2|H] = f(L(E_1, H))$.
(2) $\text{Ev}[E_1, E_2|H] = f(L(E_1, H) + L(E_2, H))$. Then inferential updating has the following form:

$$p(H|E) = \frac{e^{-c*L(E,H)}p(H)}{\sum_i e^{-c*L(E,H_i)}p(H_i)},$$

for some constant $c$.

A sketch of the proof, which is straightforward, is given in the appendix. Although general Bayesian updating is a special case of inferential updating, the reverse is not the case because—as was previously mentioned—many reasonable evidential measures cannot be written as a function of an additive loss function. Suppose, for example, that the hypotheses under consideration are real-valued functions, $f_i$ and that the evidential measure is of the form $\text{Ev}[(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)|f_i] = \text{Minimum}(|y_1 - f_i(x_1)|, |y_1 - f_i(x_1)|, \ldots, |y_1 - f_i(x_1)|)$. It is clear in this case that the evidential measure cannot be written as a function of an additive loss function, simply because the Minimum operator is not additive.
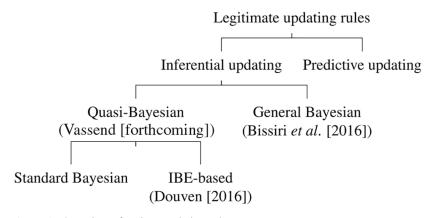
**Figure 1.** Overview of various updating rules.

A diagram depicting the relationship between inferential updating, predictive updating, and various updating rules that have been suggested in the literature is given in Figure 1.

## 5. Conclusion

The primary purpose of this article has been to justify a set of very general synchronic and diachronic inductive norms. The resulting normative framework can be put to both philosophical and scientific use. In philosophy of science, a standard way of analysing scientific methodology is by seeing whether the methodology makes sense from a Bayesian perspective. For example, in this way, Sober ([2015]) analyses parsimony inference,[12] Dawid, Hartmann and Sprenger ([2015]) analyse no-alternatives arguments in physics, Schupbach ([2018]) analyses robustness analysis, and Myrvold ([2017]) evaluates the epistemic value of unification. Since the preceding analyses take place in a Bayesian framework, they inherit the limitations and assumptions of Bayesianism. In the broader normative framework developed in this article, it's possible to check whether the analyses still hold up when those assumptions are lifted. For example, Myrvold ([2017]) shows that more unifying hypotheses will be more confirmed by evidence than less unifying hypotheses, other things being equal. Since his analysis is Bayesian, he implicitly uses the likelihood as his measure of evidential favouring. A natural question to ask is whether his result still holds if the likelihood is replaced with an arbitrary measure of evidential favouring. The perhaps surprising answer is yes, although a proper demonstration of this fact must be reserved for a different time.

The normative framework developed in this article can also be used for scientific inference. Indeed, implicitly it already has been—as shown in Section 4.2, the general

---

[12] Sober uses a likelihoodist approach, which is Bayesianism without the priors.

Bayesian updating rule suggested by Bissiri *et al.* ([2016]) is a special case of inferential updating, and general Bayesian updating is gaining in popularity in the statistical community. But inferential updating is more general than general Bayesian updating, and allows for the use of evidential measures that cannot be represented in Bissiri *et al.*'s ([2016]) framework. One example is the phylogenetic parsimony measure discussed in (Vassend ([forthcoming]). Predictive updating can also be applied in scientific inference problems, for example through the use of strictly proper scoring rules as suggested in Section 4.1. Of course, it is ultimately an empirical question whether predictive updating performs better than inferential updating. An answer to this question must wait until later; in this article, my goal has been to provide a general normative framework for inductive inference that is as flexible as possible while obeying basic theoretical desiderata.

## Acknowledgements

## Appendix

### A1. Characterization of the combination function

The goal of this section is to show the characterization of the combination function in Section 3.1. There are two cases to consider: $k = 0$ and $k \neq 0$. Since the two cases are very similar, I will only consider the case where $k \neq 0$. So suppose that for some non-zero $k$, we have:

$$\frac{\partial^2 c(x, y)}{\partial x \partial y} = k. \tag{4}$$

Taking the antiderivative with respect to $x$, it follows that

$$\frac{\partial c(x, y)}{\partial y} = kx + C(y) + D, \tag{5}$$

where $C(y)$ is a function of $y$, but not $x$, and $D$ is some real number. Taking the antiderivative of Equation (5) with respect to $y$, we get:

$$c(x, y) = kxy + \int C(y)dy + Dy + G(x) + F. \tag{6}$$

Where $G$ is a function of $x$ and $F$ is some real number. Moreover, exchanging the labels $x$ and $y$ in Equation (6) gives us:

$$c(y, x) = kyx + \int C(x)dx + Dx + G(y) + F. \tag{7}$$

But since $c(x, y) = c(y, x)$, Equation (6) and Equation (8) must be equal, which means that $kxy + \int C(y)dy + Dy + G(x) + F = kxy + \int C(x)dx + Dx + G(y) + F$, and hence $\int C(y)dy + Dy + G(x) = \int C(x)dx + Dx + G(y)$. Rearranging, we get:

$$G(x) = \int C(x)dx + Dx + G(y) - \int C(y)dy - Dy. \tag{8}$$

But since $G(x)$ does not depend on $y$, the only way for Equation (8) to be true is if $G(y) - \int C(y)dy - Dy$ is equal to some constant number, $c$. Hence, $\int C(y)dy + Dy = G(y) - c$. Plugging this back into Equation (6) (and absorbing the constant $c$ into $F$), we get:

$$c(x, y) = kxy + G(x) + G(y) + F. \tag{9}$$

Without loss of generality, we may assume that $G(0) = 0$, because if $G(0) = A$ for some non-zero $A$, then we can just put $G'(x) = G(x) - A$ and $F' = F + 2A$, and we get: $c(x, y) = kxy + G'(x) + G'(y) + F'$, with $G'(0) = 0$ (that is, we simply absorb the constant $A$ into $F'$).

Now the fact that $c$ is associative and commutative means that $c(c(x, y), z)) = c(c(y, z), x)$, and hence Equation (9) implies that for all $x$, $y$, and $z$:

$$k(kxy + G(x) + G(y) + F)z + G(kxy + G(x) + G(y) + F) + G(z) + F$$
$$= k(kyz + G(y) + G(z) + F)x + G(kyz + G(y) + G(z) + F) + G(x) + F \tag{10}$$

Simplifying, we have:

$$[G(x) + G(y) + F]kz + G[kxy + G(x) + G(y) + F] + G(z)$$
$$= G(y)kx + G(z)kx + Fkx + G[kyz + G(y) + G(z) + F] + G(x) \tag{11}$$

Note that because $c$ is twice differentiable, so is $G$. Taking the derivative of each side of Equation (11) with respect to $z$ gives:

$$[G(x) + G(y) + F]k + \frac{\partial G(z)}{\partial z} = \frac{\partial G(z)}{\partial z}kx + G'[kyz + G(y) + G(z) + F] * \frac{\partial G(z)}{\partial z}. \tag{12}$$

Next, taking the derivative of each side of Equation (12) with respect to $x$ gives:

$$\frac{\partial G(x)}{\partial x}k = \frac{\partial G(z)}{\partial z}k. \tag{13}$$

a multiplicative normalization step, or (2) an additive combination step and an additive normalization step. It is easy to show that it is possible for an updating rule that satisfies (1) to satisfy regularity, and that—indeed—the resulting updating rule is inferential updating. In order to show that inferential updating is the only updating rule that satisfies regularity, it suffices to show that there is no updating rule satisfying (2) that also satisfies regularity.

Suppose, for the sake of contradiction, that there is some updating rule that satisfies both (2) and regularity. In order for regularity to be obeyed, it has to be the case that given any set of non-zero prior probabilities over a set of hypotheses, $h_1, h_2, \ldots, h_n$, and given any set of evidential scores for the hypotheses, $e_1, e_2, \ldots, e_n$, the posteriors are also all non-zero. Thus, if $N$ is the normalization function, then the following must be true for all $h_i$:

$$N(e_i + h_i) > 0. \tag{16}$$

Since the normalization function is assumed to satisfy (2), Equation (16) implies that the following is true for all $i$, where $d$ is an additive normalization constant:

$$e_i + h_i + d > 0. \tag{17}$$

Since the posterior probabilities must sum to one, we also have:

$$\sum_i (e_i + h_i + d) = 1. \tag{18}$$

And therefore, $d = -\frac{1}{n}\Sigma e_i$. And so we have, for all $h_i$:

$$e_i + h_i - \frac{1}{n}\sum e_i > 0. \tag{19}$$

But it's obvious that Equation (19) will not in general be true. For example, suppose $e_1$ is the smallest $e_i$. Then $r = e_i - \frac{1}{n}\Sigma e_i < 0$. Now suppose it's also the case that $h_1 < -r$. Then we have:

$$e_1 + h_1 - \frac{1}{n}\sum e_i = r + h_1 < 0. \tag{20}$$

Consequently, additive combination and additive normalization jointly violate regularity. So there can be no updating procedure that satisfies both (2) and regularity.

## A4. Characterization of predictive updating

The goal in this section is to show that the only legitimate updating rule that violates regularity but satisfies conservativeness is predictive updating. It is clear that any updating rule that satisfies conservativeness but violates regularity must be additive. This is because any multiplicative updating rule that satisfies conservativeness clearly also satisfies regularity.

So suppose the updating rule is additive and satisfies conservativeness. Then the goal is to show that the updating rule must be equivalent to predictive updating. Since the rule is additive, it must have the following form, where $p_E$ is the posterior probability distribution, $H_i$ is a hypothesis, $h_i$ is the prior probability of the hypothesis, $e_i$ is the evidential score of the hypothesis, and $d$ is a normalization constant:

$$p_E(H) = \begin{cases} 0 & \text{given that } x \text{ is sufficiently low;} \\ h_i + e_i + d & \text{otherwise.} \end{cases} \tag{21}$$

If the updating rule is conservative, then as few hypotheses as possible should be assigned a posterior probability of zero. It remains to show that this uniquely happens when $d$ is minimal. Suppose there are $n$ hypotheses. Without loss of generality, suppose the hypotheses are ordered such that $0 \geq p_E(H_1) \geq p_E(H_2) \geq \ldots \geq p_E(H_n)$. Then there is some index $m$ such that $p_E(H_i) = 0$ for $i \leq m$ and $p_E(H_i) > 0$ for $i > m$. Note that the updating procedure is conservative if and only if $m$ is minimal because $m$ is minimal if and only if a minimal number of hypotheses have a posterior probability of zero. In order for the posterior probabilities to be probabilistic, we must have:

$$\sum_i p_E(H_i) = \sum_{i>m}(h_i + e_i) + (n - m)d = 1. \tag{22}$$

Now suppose we have a different updating rule resulting in some posterior $p'$ that is not conservative, so that there is an index $m' > m$ such that $p'_E(H_i) = 0$ for $i \leq m'$ and $p'_E(H_i) > 0$ for $i > m'$. Then $p'$ must satisfy the following constraint for some normalization constant $d'$:

$$\sum_{i>m'}(h_i + e_i) + (n - m')d' = 1. \tag{23}$$

Comparing Equation (21) and Equation (23) and remembering that $m' > m$, we see that:

$$0 < \sum_{i=m}^{m'}(h_i + e_i) = (n - m')d' - (n - m)d. \tag{24}$$

And hence,

$$d < \frac{n - m'}{n - m}d' < d'. \tag{25}$$

Hence, $d < d'$. What the above proof shows is that any conservative updating rule has a smaller additive normalization constant than any non-conservative updating rule. To finish the proof, we show that there is just one conservative updating rule. Here we can use Equation (24) again. If both updating rules are conservative, then we have $m = m'$, and hence—making the necessary amendments in Equation (24)—we have:

$$0 = \sum_{i=m}^{m'}(h_i + e_i) = (n - m)d' - (n - m)d. \tag{26}$$

Hence it follows that $d' = d$. But then the two updating rules are equivalent. Hence, there is only one conservative updating rule, namely, the one that uses a minimal additive normalization constant. This is predictive updating.

## A5. General Bayesian updating is a special case of inferential updating

The goal in this section is to show that Bissiri *et al.'s* ([2016]) general Bayesian updating is a special case of inferential updating. For some normalization constant $k$, we have:

$$p(H|E_1, E_2) = k * \text{Ev}[E_1|H, E_2]\text{Ev}[E_2|H]p(H)$$
$$= k * f(L(E_1, H))f(L(E_2, H))p(H). \tag{27}$$

But we also have:

$$p(H|E_1, E_2) = k * \text{Ev}[E_1, E_2|H]p(H) = k * f(L(E_1, H) + L(E_2, H))p(H). \tag{28}$$

Comparing Equation (27) and Equation (28), we see that $f$ obeys the following functional equation for all $x$ and $y$: $f(x)f(y) = f(x + y)$. Let $g(x) = \log f(x)$. Then $g(x + y) = g(x) + g(y)$, which is the well-known Cauchy equation whose solution is $g(x) = -cx$, for some positive constant $c$ (Aczél [2006], p. 31) (since $f$, and therefore $g$, is strictly decreasing). Consequently, $f(x) = e^{-cx}$, and hence $p(H|E) = k * e^{-c*L(E,H)}p(H)$, which is Bissiri *et al.*'s ([2016]) general Bayesian updating rule.

## A6. An alternative characterization of the combination step

In both everyday and scientific contexts, it's common to think of evidence algebraically: multiple lines of evidence combine in order provide stronger evidence; some evidence favours a hypothesis, while other evidence goes against it; a piece of evidence here can cancel out a piece of evidence there; and some purported evidence has no effect at all. In other words, evidential favouring has all the hallmarks of a mathematical group. Now, suppose—as we have been doing up to now—that we use real numbers to represent evidential scores. Then the set of all possible evidential scores, $G$, together with the combination function plausibly form a mathematical group. Indeed, they plausibly form an Archimedean group, because intuitively there is no maximal evidential score. That is, if we use • to denote the combination function, so that $e_1 • e_2 = c(e_1, e_2)$, then it is plausible that $(G, •)$ satisfies the following axioms:

> Closure: For all possible evidential scores $e_1$ and $e_2$, $e_1 • e_2$ is also a possible evidential score.

> Associativity: For all possible evidential scores $e_1$, $e_2$ and $e_3$, $(e_1 • e_2) • e_3 = e_1 •(e_2 • e_3)$.

Identity: There exists a possible evidential score $i$ such that for all $e$, $i \bullet e = e \bullet i = e$. That is, there exists a real number that represents evidence that has no effect (either favourable or unfavourable).

Inverse: For each possible evidential score $e$, there exists a possible evidential score $e'$ such that $e \bullet e' = e' \bullet e = i$. That is, every evidential score could potentially (in principle) be cancelled out by other countervailing evidence.[14]

Commutativity: For all possible evidential scores $e_1$ and $e_2$, $e_1 \bullet e_2 = e_2 \bullet e_1$. That is, the order in which the evidence is considered is irrelevant.

Archimedean Property: For all possible evidential scores $e_1$ and $e_2$, there exists an integer $n$ such that $e_1 < e_2 \bullet e_2 \ldots \bullet e_2$ ($n$ times).

Suppose, in addition, that the set of evidential scores is totally ordered: for all evidential scores $e_1$ and $e_2$, either $e_1 > e_2$ or $e_1 \leq e_2$.[15] Then we can use the following important result from group theory (see p. 33 of Kopytov and Medvedev [1996] for a proof):

Hölder's Theorem: Every Archimedean totally ordered group is order-isomorphic to a subgroup of the additive group of real numbers with the natural order.

The fact that $(G, \bullet)$ is order-isomorphic to a subgroup of the additive group of real numbers with the natural order means there exists some subgroup, $(S, +)$ of the real numbers and a one-to-one function, $g$, from $(G, \bullet)$ to $(S, +)$ that obeys the following equation for all $e_1$ and $e_2$ in $G$: $g(e_1 \bullet e_2) = g(e_1) + g(e_2)$. Since $g$ is one-to-one, it has an inverse, $f$. Hence, for all $e_1$ and $e_2$ in $G$, we can write: $e_1 \bullet e_2 = f(g(e_1) + g(e_2))$.

In the main text, I showed that the normalization procedure must be either additive or multiplicative, given that the combination function is either multiplicative or additive. But, arguably, it is not unreasonable to simply assume that the normalization must be either multiplicative or additive. Indeed, all updating rules that have been proposed in the literature have implicitly relied on a normalization procedure that is either multiplicative or additive. In particular, the normalization procedure implicit in both standard Bayesian updating and Jeffrey updating (Jeffrey [1983]) is multiplicative, and the normalization procedure implicit in Leitgeb and Pettigrew's ([2010]) alternative to Jeffrey updating is additive.

Finally, it is reasonable to assume—as we did in the main text—that the normalization procedure commutes with the combination function in the sense that for all $a$, $b$, and $c$, we have: $N(a \bullet N(b)) = N(N(a) \bullet b) = N(a \bullet b)$. We can now give the following characterization of the combination function:

Alternative Characterization of the Combination Function: Suppose the combination function, $c(x, y)$ satisfies the following requirements:

---

[14] A referee points out that this is a bit of an idealization, since a piece of evidence and a defeater of that evidence will not typically cancel each other out precisely.
[15] A referee rightly points out that this assumption is also idealized.

(1) The set of all evidential scores, $G$, and the combination function $c(x, y) = x \bullet y$ together form a totally ordered Archimedean group.

(2) The combination function commutes with the normalization function $N$ in the sense that for all $a$, $b$, and $c$: $N(a \bullet N(b)) = N(N(a) \bullet b) = N(a \bullet b)$. Then $c$ must have one of the following two forms:

    (a) If the normalization function is additive, then $c(x, y) = x + y$.

    (b) If the normalization function is multiplicative, then $c(x, y) = xy$.

## Proof

The fact that the combination function commutes with the normalization function implies that for every $e$ with inverse $e^{-1}$:

$$N(e \bullet e^{-1}) = N(N(e) \bullet e^{-1}) = N(f(g(N(e)) + g(e^{-1}))). \qquad (29)$$

Therefore, for all $e$, $N(f(g(N(e)) + g(e^{-1}))) = N(i)$, where $i$ is the identity element of the group. Since $N$ is one-to-one, this means that $f(g(N(e)) + g(e^{-1})) = k$, for some constant $k$ that does not depend on $e$. Furthermore, since $f$ is one-to-one, this in turn implies that $g(N(e)) + g(e^{-1}) = k'$, for some constant $k'$ that does not depend on $e$. For the same reason, Equation (29) also implies that $g(e) + g(e^{-1}) = k''$, for some constant $k''$ that does not depend on $e$. Hence we have, finally, that $g(N(e)) - g(e) = K$, where $K = k' - k''$. Hence, $g(N(e)) = g(e) + K$.

If the normalization procedure is multiplicative, then for some normalization constant $a$, we have $g(ae) = g(e) + K$. Note that $a$ depends on the set to which $e$ belongs. If $\{e_i\}$ is the set, then

$$a = \frac{1}{\sum e_i}. \qquad (30)$$

Hence, depending on the other members of the set to which $e$ belongs, $a$ can be any number in the half-open interval $(0, \frac{1}{e})$. Thus we have, for all $e$ and all $a$ in $(0, \frac{1}{e})$, that $g(ae) = g(e) + K$, where $K$ is a constant that may depend on $a$, but does not depend on $e$.

Similarly, we have—for some normalization constant $b$—that $g(bae) = g(ae) + K' = g(e) + K''$. Here, $b$ can be any number in the range $(0, \frac{1}{ae})$, or in other words in $(0, \infty)$. But if we let $y = ab$ and $x = e$, then the preceding means that for all $x$ and $y$ in $(0, \infty)$ we have:

$$g(yx) = g(x) + K'', \qquad (31)$$

where $K''$ depends on $y$, but not on $x$. Interchanging the role of $y$ and $x$, we also have:

$$g(xy) = g(y) + K''', \qquad (32)$$

where $K'''$ depends on $x$, but not on $y$. Comparing the above equations, we see that $g(x) + K'' = g(y) + K'''$. This implies the following:

$$g(xy) = g(x) + g(y) + C, \qquad (33)$$

where $C$ is a constant that depends on neither $x$ nor $y$. Now note that $f(2g(i)) = i \bullet i = i = f(g(i))$. Since $f$ is one-to-one, this implies that $g(i) = 0$. Next, Equation (33) implies that $g(i) = g(1 * i) = g(1) + g(i) + C$. Thus $g(1) = -C$. Using Equation (33) again, we have $g(1) = g(i * \frac{1}{i}) = g(i) + g(\frac{1}{i}) = g(\frac{1}{i})$. But since $g$ is one-to-one, this implies that $\frac{1}{i} = 1$, so that $i = 1$. Hence $-C = g(1) = g(i) = 0$, so $C = 0$. Finally, then, we have, for all $x > 0$ and $y > 0$:

$$g(xy) = g(x) + g(y). \qquad (34)$$

Now put $r(x) = g(e^x)$. Then Equation (34) becomes, for all real $x$ and $y$:

$$r(x + y) = r(x) + r(y). \qquad (35)$$

This is the Cauchy functional equation, whose only solution is $r(x) = cx$, for an arbitrary constant $c$ (Aczél [2006], p. 31). Hence, $g(x) = r(\log x) = \log x^c$. Since $f$ is the inverse of $g$, we have that $f(x) = e^{x^{\frac{1}{c}}}$. Finally, then, we have:

$$x \bullet y = f(g(x) + g(y)) = e^{(\log(x^c) + \log(y^c))^{\frac{1}{c}}} = e^{(c * \log(xy))^{\frac{1}{c}}} = xy. \qquad (36)$$

Thus the combination function is multiplicative, $c(x, y) = xy$. □

*Philosophy Programme*
*Nanyang Technological University*
*Singapore, Singapore*
*vassend@ntu.edu.sg*

# References

Aczél, J. [2006]: *Lectures on Functional Equations and Their Applications*, Mineola, NY: Dover Publications.

Amari, S.-I. [2009]: 'Alpha-Divergence Is Unique, Belonging to Both f-Divergence and Bregman Divergence Classes', *IEEE Transactions on Information Theory*, **55**, pp. 4925–31.

Bernardo, J. M. and Smith, A. F. M. [1994]: *Bayesian Theory*, New York: Wiley.

Bissiri, P. G., Holmes, C. and Walker, S. [2016]: 'A General Framework for Updating Belief Distributions', *Journal of the Royal Statistical Society: Series B*, **78**, pp. 1103–30.

Box, G. E. P. [1980]: 'Sampling and Bayes' Inference in Scientific Modelling and Robustness', *Journal of the Royal Statistical Society: Series A (General)*, **143**, pp. 383–430.

Dawid, R., Hartmann, S. and Sprenger, J. [2015]: 'The No Alternatives Argument', *British Journal for the Philosophy of Science*, **66**, pp. 213–34.

Douven, I. [2016]: 'Explanation, Updating, and Accuracy', *Journal of Cognitive Psychology*, **28**, pp. 1004–12.

Douven, I. and Wenmackers, S. [2017]: 'Inference to the Best Explanation versus Bayes's Rule in a Social Setting', *British Journal for the Philosophy of Science*, **68**, pp. 535–70.

Forster, M. R. [1995]: 'Bayes and Bust: Simplicity as a Problem for a Probabilist's Approach to Confirmation', *British Journal for the Philosophy of Science*, **46**, pp. 399–424.

Forster, M. R. and Sober, E. [1994]: 'How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions', *British Journal for the Philosophy of Science*, **45**, pp. 1–35.

Gelman, A. and Shalizi, C. R. [2013]: 'Philosophy and the Practice of Bayesian Statistics', *British Journal of Mathematical and Statistical Psychology*, **66**, pp. 8–38.

Gneiting, T. and Raftery, A. E. [2007]: 'Strictly Proper Scoring Rules, Prediction, and Estimation', *Journal of the American Statistical Association*, **102**, pp. 359–78.

Greaves, H. and Wallace, D. [2006]: 'Justifying Conditionalization: Conditionalization Maximizes Epistemic Utility', *Mind*, **115**, pp. 607–32.

Grünwald, P. and van Ommen, T. [2017]: 'Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It', *Bayesian Analysis*, **12**, pp. 1069–103.

Jeffrey, R. [1983]: *The Logic of Decision*, Cambridge: Cambridge University Press.

Joyce, J. [1998]: 'A Non-pragmatic Vindication of Probabilism', *Philosophy of Science*, **65**, pp. 575–603.

Joyce, J. [2009]: 'Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief', in F. Huber and C. Schmidt-Petri (*eds*), *Degrees of Belief*, Dordrecht: Springer.

Key, J. T., Pericchi, L. R. and Smith, A. F. M. [1999]: 'Bayesian Model Choice: What and Why?', in J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (*eds*), *Bayesian Statistics 6*, Oxford: Oxford University Press, pp. 343–70.

Kopytov, V. M. and Medvedev, N. Y. [1996]: *Right-Ordered Groups*, New York: Springer.

Leitgeb, H. and Pettigrew, R. [2010]: 'An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy', *Philosophy of Science*, **77**, pp. 236–72.

Levinstein, B. A. [2012]: 'Leitgeb and Pettigrew on Accuracy and Updating', *Philosophy of Science*, **79**, pp. 413–24.

Myrvold, W. [2017]: 'On the Evidential Import of Unification', *Philosophy of Science*, **84**, pp. 92–114.

Pettigrew, R. [2016]: *Accuracy and the Laws of Credence*, Oxford: Oxford University Press.

Predd, J. B., Seiringer, R., Lieb, E. H., Osherson, D. N., Poor, H. V. and Kulkarni, S. R. [2009]: 'Probabilistic Coherence and Proper Scoring Rules', *IEEE Transactions on Information Theory*, **55**, pp. 4786–92.

Schupbach, J. N. [2018]: 'Robustness Analysis as Explanatory Reasoning', *British Journal for the Philosophy of Science*, **69**, pp. 275–300.

Shaffer, M. J. [2001]: 'Bayesian Confirmation of Theories That Incorporate Idealizations', *Philosophy of Science*, **68**, pp. 36–52.

Sober, E. [2015]: *Ockham's Razors: A User's Manual*, Cambridge: Cambridge University Press.

Sprenger, J. [2009]: 'Statistics between Inductive Logic and Empirical Science', *Journal of Applied Logic*, **7**, pp. 239–50.

Sprenger, J. [forthcoming]: 'Conditional Degree of Belief', *Philosophy of Science*.

Tanton, J. S. [2005]: *Encyclopedia of Mathematics*. New York: Facts of File.

Vassend, O. B. [2019]: 'New Semantics for Bayesian Inference: The Interpretive Problem and Its Solutions', *Philosophy of Science*, **86**, pp. 696–718.

Vassend, O. B. [forthcoming]: 'A Verisimilitude Framework for Inductive Inference, with an Application to Phylogenetics', *British Journal for the Philosophy of Science*, available at https://doi.org/10.1093/bjps/axy054.

Vassend, O. B. [unpublished]: 'The Formal Structure of Epistemic Trust', available at <sites .google.com/site/olavbvassend/research>.

Walker, S. G. [2013]: 'Bayesian Inference with Misspecified Models', *Journal of Statistical Planning and Inference*, **143**, pp. 1621–33.

Zhang, T. [2006]: 'From e-Entropy to KL-Entropy: Analysis of Minimum Information Complexity Density Estimation', *The Annals of Statistics*, **34**, pp. 2180–210.