# Justifying the Norms of Inductive Inference

January 14, 2019

**Abstract**

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 favoring, even when that is inappropriate. The purpose of this paper 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 favoring 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, and Vassend's (2019a) quasi-Bayesian updating as special cases. The other updating rule is novel.

## Contents

# 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 favoring: if $H_1$'s likelihood on the evidence is greater than $H_2$'s likelihood on the same evidence, then the evidence favors $H_1$ 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 favoring. 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 Vassend (2019a) shows, the likelihood is not an appropriate measure of evidential favoring 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 favoring is one according to which the evidence favors $H_1$ 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 favoring is therefore a limitation of the framework.

The goal of this paper is to study inductive inference in a very general setting. Suppose our goal is to identity the best hypothesis $H$ (where "best" does not necessarily mean "true"). Let $p$ be function that assigns a number between 0 and 1 (inclusive) to each hypothesis, such that $p(H)$ is interpreted as representing a prior judgment of how plausible it is that $H$ is best (in the relevant sense) out of the hypotheses under consideration. 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

[1]This limitation is well known, but often ignored. For discussion of the problem, see, e.g. Box (1980); Bernardo and Smith (1994); Forster and Sober (1994); Forster (1995); Key et al. (1999); Shaffer (2001); Sprenger (2009); Gelman and Shalizi (2013); Vassend (2019b); Walker (2013); and Sprenger (2017).

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 $p$ 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:
>
> > **Step 1.** For each $i$, calculate $q(H_i) = p(H_i) + \mathrm{Ev}[E|H_i]$.
> >
> > **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 $\sum_i p_E(H_i) = 1$.

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 favoring. What separates inferential updating from predictive updating is the former rule's commitment to *Regularity*: inferential updating will never assign a probability of 0 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 paper is as follows. In Section 2, I sketch an argument for why any plausibility 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 favoring 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 paper with a discussion of inferential and predictive updating, including their relationship to each other and to other updating rules.

# 2 Why plausibility functions should be probabilistic

One of the standard arguments for why regular plausibilities (or degrees of belief) ought to be probabilistic is the accuracy argument (Joyce (1998), Joyce (2009), Pettigrew (2016), Predd et al. (2009)). Briefly, the argument is as follows:[2] the ideal plausibility function to have is the function that assigns 1 to the hypothesis that is true and 0 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 plausibility function. It can then be shown that any plausibility 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

[2]There are several versions of the argument; here, I present a variant of Pettigrew's (2016) version.

irrational to choose an option that is known to be dominated, it follows that it is irrational to use a non-probabilistic plausibility 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 plausibility function, nor does it depend on the assumption that the ideal plausibility function is the function that assigns 1 to the hypothesis that is true and 0 to all hypotheses that are false. Indeed, nothing in the accuracy argument prevents us from designating the ideal plausibility 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 (where "best" can mean anything we like). In such a context, an ideal function would clearly be one that assigns 1 to the hypothesis that is best and 0 to all other hypotheses. 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] For example, if "best" means "having a minimal maximum expected prediction error," then there may be several hypotheses that are tied for best. However, it is easy to accommodate this complication, as I have done in the following generalization of the accuracy argument:

> P1: An ideal plausibility function is any function that assigns 1 to a hypothesis that is best and 0 to all other hypotheses.
>
> P2: For any ideal plausibility function and 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: For any ideal plausibility function and 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 plausibility functions are irrational.

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

C: Non-probabilistic plausibility 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 paper is not to give a careful analysis of the accuracy argument. From now I will assume that any plausibility function ought to be probabilistic. That is, I will assume that if $p$ is a function that assigns a number between 0 and 1 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 paper: 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

It is widely accepted that if the goal is to find the true hypothesis in a partition of hypotheses, then any probability function over the hypotheses ought to be updated through Bayesian updating:

**Bayesian updating:** $p(H|E) = \frac{p(E|H)p(H)}{\sum_i p(E|H_i)p(H_i)}$

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 plausibility 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 (e.g. 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 $\mathrm{Ev}[E_2\&E_2|H] = f(L(E_1, H) + L(E_2, H))$, and given that a few other assumptions are met, then the updating procedure must have the following form, where $c$ is some constant:

$$p(H|E) = \frac{e^{-c*L(E|H)}p(H)}{\sum_i e^{-c*L(E|H_i)}p(H_i)} \tag{3.1}$$

Bissiri et al. (2016) call the above updating procedure "general Bayesian updating." General Bayesian updating was originally introduced (not under that name) by Zhang (2006) and has been increasingly influential in statistics in recent years. Although Bissiri et al.'s (2016) argument for general Bayesian updating is interesting, it has several limitations. One problem is that, as Vassend (2019b) argues, the probabilities in (3.1) cannot be interpreted in the standard Bayesian way as plausibilities of truth. But if the probabilities are not standard plausibility 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. This assumption rules out many standard divergence measures, including all Bregman divergences aside from

the Kullback-Leibler divergence (Amari, 2009).[4] 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 $i$.[5] 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 (i.e. likelihood) of the hypothesis in order to produce an overall judgment 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, i.e. such that all the posterior plausibility scores fall between 0 and 1, inclusive, and jointly sum to 1.

It is reasonable to suppose that any updating procedure may be similarly decomposed into a combination step and a normalization step. The combination step requires a combination function, $c$, that takes as its input a prior probability, $p(H)$ and a set of evidential scores, $\mathrm{Ev}[E_1|H]$, $\mathrm{Ev}[E_2|H, E_1]$, $\mathrm{Ev}[E_3|H, E_1, E_2]$, etc., 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

[4]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)).

[5]If $p(E_1, E_2|H) = p(E_1|H)p(E_2|H)$, we can write $p(E_1, E_2|H) = e^{\log p(E_1|H) + \log p(E_2|H)}$, i.e. 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.

those scores into probabilities. In other words, on an abstract level, any updating procedure may plausibly be decomposed 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 analyzed 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

There really are only two plausible candidate forms for the combination function: either the function is additive or it is multiplicative. That is, 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 the combination function plausibly has one of the following forms:

> **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 $\mathrm{Ev}[E_1|H]$, i.e. the unconditional evidential score of $H$ on $E_1$, or it may represent $\mathrm{Ev}[E_1|H, E_2]$, i.e. 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 $\mathrm{Ev}[E_1, E_2|H] = \mathrm{Ev}[E_1|H] + \mathrm{Ev}[E_2|H]$ or $\mathrm{Ev}[E_1, E_2|H] = \mathrm{Ev}[E_1|H] * \mathrm{Ev}[E_2|H]$. 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 judgment should not influence the overall judgment 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 $\mathrm{Ev}[E_1|H, E_2]$. 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, $e_1 = \mathrm{Ev}[E_1|H, E_2]$ makes on $H$'s total evidential score should not depend on the impact that $e_2 = \mathrm{Ev}[E_2|H]$ makes on $H$'s total evidential score, nor vice versa.

The preceding requirement may be naturally formalized as constraints on the partial derivatives of the combination function. Let $c(x, y)$ be the combination func-

tion 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 0, 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, i.e. $\frac{\partial c}{\partial x}$, should not depend on the impact that $y$ makes, i.e. $\frac{\partial c}{\partial y}$, where $x$ and $y$ represent any possible evidential scores. We can now show the following (the derivation is in Appendix A):

> **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. $c$'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 Appendix F.

## 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 0 and 1 and jointly sum to 1, 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 0, while the remaining numbers in the set are rescaled by some function, $f$. In other words, normalization in general takes the following functional form:

$$N(x) = \begin{cases} 0 & \text{Given that } x \text{ is sufficiently low} \\ f(x) & \text{Otherwise} \end{cases} \tag{3.2}$$

For example, in the normalization step of standard Bayesian updating, $N(x) = f(x)$ (i.e. no non-zero numbers are normalized to 0) 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 0, 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 judgment. 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 (3.2) 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 Appendix B.

## 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. 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, i.e. the posterior probability of any hypothesis is always greater than 0.

We can then show the following (see Appendix C):

**Characterization of inferential updating.** The only legitimate updating procedure that satisfies Regularity is inferential updating. I.e., given evidential measure Ev and prior probability function $p$, update $p$ to the posterior $p_E$ by way of the following formula:

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

Inferential updating satisfies Regularity; it will never result in any hypothesis having a posterior probability of 0. On the other hand, in Appendix C, I show that an updating procedure that uses an additive combination function and an additive normalization function must violate Regularity; most of the time, any such updating rule must assign a posterior probability of 0 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—i.e., assign them a posterior probability of 0.

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 0 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. I.e., 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 $q(H_i) = p(H_i) + \mathrm{Ev}[E|H_i]$.
>
> **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 $\sum_i p_E(H_i) = 1$.

# 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, i.e. 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 0 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 favoring (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 $M_1$, $M_2$, etc., and for each model, let $p_{M_i}$ be the marginal probability forecast distribution corresponding to $M_i$. Suppose, moreover, that $p_{M_i}$ has finite first moment. Then the CRPS can be written in the following way (where the expectations are taken relative to $p_{M_i}$):

$$\mathrm{CRPS}(p_{M_i}, x) = \mathrm{E}|X - x| - \frac{1}{2}\mathrm{E}|X_1 - X_2| \tag{4.1}$$

As (4.1) 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 (4.1) 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 will be a representation of how many meters inaccurate the model's predictions are, on average.

If we let $\mathrm{Ev}[x|p_{M_i}] = a * \mathrm{CRPS}(p_{M_i}, x)$, 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.[6] Importantly, given sufficient evidence (and depending on how the constant $a$ is chosen) many of the models will receive a posterior probability of 0. 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 paper, standard Bayesian updating is clearly a special case of inferential updating: more precisely, we get Bayesian updating if and only if $\mathrm{Ev}[E|H] \propto p(E|H)$, i.e. if and only if the evidential measure is proportional to the likelihood. What Vassend (2019a) 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.

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. $\mathrm{Ev}[E_1|H, E_2] = \mathrm{Ev}[E_2|H] = f(L(E_1, H))$.

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

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 Appendix E. 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.

# 5 Conclusion

The primary purpose of this paper 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 analyzing scientific methodology is by seeing whether the methodology makes sense from a Bayesian perspective. For example, in this way, Sober (2015) analyzes parsimony inference,[7] Dawid et al. (2015) analyze no-alternatives arguments in physics, Schupbach (2018) analyzes robustness analysis, and Myrvold (2016) 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 paper, it's possible to check whether

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

the analyses still hold up when those assumptions are lifted. For example, Myrvold (2016) 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 favoring. A natural question to ask is whether his result still holds if the likelihood is replaced with an arbitrary measure of evidential favoring. 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 paper can also be used for scientific inference. Indeed, implicitly it already has been—as shown in Section 4.2, the general 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 by Vassend (2019a). 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 paper, my goal has been to provide a general normative framework for inductive inference that is as flexible as possible while obeying basic theoretical desiderata.

# References

Aczél, J. (2006). *Lectures on Functional Equations and Their Applications*. Dover Books on Mathematics. Dover Publications.

Amari, S.-I. (2009). α-Divergence is Unique, Belonging to Both $f$-Divergence and Bregman Divergence Classes. *IEEE Transactions on Information Theory 55*(11), 4925 – 4931.

Bernardo, J. M. and A. F. M. Smith (1994). *Bayesian Theory*. Wiley, New York, NY.

Bissiri, P. G., C. Holmes, and S. Walker (2016). A General Framework for Updating Belief Distributions. *Journal of the Royal Statistical Society. Series B (Methodological) 78*(5), 1103–1130.

Box, G. E. P. (1980). Sampling and Bayes' Inference in Scientific Modelling and Robustness. *Journal of the Royal Statistical Society. Series A (General) 143*(4), 383–430.

Dawid, R., S. Hartmann, and J. Sprenger (2015). The No Alternatives Argument. *British Journal for the Philosophy of Science 66*(1), 213–234.

Forster, M. R. (1995, September). Bayes and bust: Simplicity as a problem for a probabilist's approach to confirmation. *British Journal for the Philosophy of Science 46*(3), 399–424.

Forster, M. R. and E. Sober (1994). How To Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions. *The British Journal for the Philosophy of Science 45*(1), 1–35.

Gelman, A. and C. R. Shalizi (2013). Philosophy and the Practice of Bayesian Statistics. *British Journal of Mathematical and Statistical Psychology 66*, 8–38.

Gneiting, T. and A. E. Raftery (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association 102*(477), 359–378.

Greaves, H. and D. Wallace (2006). Justifying conditionalization: Conditionalization maximizes epistemic utility. *Mind 115*(459), 607–632.

Jeffrey, R. (1983). *The Logic of Decision* (Second ed.). Cambridge University Press, Cambridge.

Joyce, J. (1998). A Non-Pragmatic Vindication of Probabilism. *Philosophy of Science 65*(4), 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*. Synthese.

Key, J. T., L. R. Pericchi, and A. F. M. Smith (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*, pp. 343–370. Oxford: Oxford University Press.

Kopytov, V. M. and N. Y. Medvedev (1996). *Right-Ordered Groups*. Siberian School of Algebra and Logic. Springer.

Leitgeb, H. and R. Pettigrew (2010). An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy. *Philosophy of Science 77*, 236–272.

Myrvold, W. (2016). On the Evidential Import of Unification. Unpublished manuscript.

Pettigrew, R. (2016). *Accuracy and the Laws of Credence*. Oxford University Press.

Predd, J. B., R. Seiringer, E. H. Lieb, D. N. Osherson, H. V. Poor, and S. R. Kulkarni (2009). Probabilistic Coherence and Proper Scoring Rules. *IEEE Transactions on Information Theory 55*(10), 4786–4792.

Schupbach, J. N. (2018). Robustness Analysis as Explanatory Reasoning. *British Journal for the Philosophy of Science 69*(1), 275–300.

Shaffer, M. J. (2001). Bayesian Confirmation of Theories That Incorporate Idealizations. *Philosophy of Science 68*(1), 36–52.

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

Sprenger, J. (2009). Statistics Between Inductive Logic and Empirical Science. *Journal of Applied Logic 7*(2), 239–250.

Sprenger, J. (2017). Conditional Degree of Belief. Unpublished manuscript.

Vassend, O. B. (2019a). A Verisimilitude Framework for Inductive Inference, with an Application to Phylogenetics. To appear in *British Journal for the Philosophy of Science*.

Vassend, O. B. (2019b). New Semantics for Bayesian Inference: The Interpretive Problem and Its Solutions. To appear in *Philosophy of Science*.

Walker, S. G. (2013). Bayesian Inference with Misspecified Models. *Journal of Statistical Planning and Inference 143*(10), 1621–1633.

Zhang, T. (2006). From e-Entropy to KL-Entropy: Analysis of Minimum Information Complexity Density Estimation. *The Annals of Statistics 34*(5), 2180–2210.

# A 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:

$$\frac{\partial^2 c(x,y)}{\partial x \partial y} = k \tag{A.1}$$

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

$$\frac{\partial c(x,y)}{\partial y} = kx + C(y) + D \tag{A.2}$$

Where $C(y)$ is a function of $y$, but not $x$, and $D$ is some real number. Taking the antiderivative of A.2 with respect to $y$, we get:

$$c(x,y) = kxy + \int C(y)dy + Dy + G(x) + F \tag{A.3}$$

Where $G$ is a function of $x$ and $F$ is some real number. Moreover, since $c(x,y) = c(y,x)$, A.3 implies that $kxy + \int C(y)dy + Dy + G(x) + F = kxy + \int C(x)dx + Dx +$

$G(y) + F$, and hence $\int C(y)dy + Dy + G(x) = \int C(x)dx + Dx + G(y)$. Comparing the terms that depend on $x$, we see that $G(x) = \int C(x)dx + Dx$. Hence, A.3 implies that:

$$c(x, y) = kxy + G(x) + G(y) + F \tag{A.4}$$

Now the fact that $c$ is associative and commutative means that $c(c(x, y), z)) = c(c(y, z), x)$, and hence A.4 implies that, for all $x$, $y$, and $z$:

$$\begin{aligned} &k(kxy + G(x) + G(y) + F)z + G(kxy + G(x) + G(y) + F) + G(z) + F \\ &= k(kyz + G(y) + G(z) + F)x + G(kyz + G(y) + G(z) + F) + G(x) + F \end{aligned} \tag{A.5}$$

Simplifying, we have:

$$\begin{aligned} &[G(x) + G(y) + F]kz + G[kxy + G(x) + G(y) + F] + G(z) \\ &= G(y)kx + G(z)kx + Fkx + G[kyz + G(y) + G(z) + F] + G(x) \end{aligned} \tag{A.6}$$

Comparing the terms in A.6 that depend on $z$, we see that:

$$[G(x) + G(y) + F]kz + G(z) = G(z)kx + G[kyz + G(y) + G(z) + F] \tag{A.7}$$

And comparing the terms in A.7 that depend on $x$, we see that $G(x)kz = G(z)kx$. Hence, $G(x) = ax$, for some constant $a$. Next, the fact that $c(x, y, z) = c(c(x, y), z)$, implies:

$$kxyz+ax+ay+az+F = k(kxy+ax+ay+F)z+a(kxy+ax+ay+F)+az+F \tag{A.8}$$

Comparing the terms that contain $xyz$, we see that $k = 1$, and hence:

$$ax + ay = axz + ayz + Fz + axy + a^2x + a^2y + Fa \tag{A.9}$$

Comparing the terms that contain $z$, we see that $a(x + y) + F = 0$ for all $x$ and $y$. The only way this can be true is if $a = F = 0$. Hence we have, finally, that $c(x, y) = xy$.

# B Characterization of the normalization step

The goal of this section is to show the characterization of the normalization step in Section 3.2. Let $\{a_i\}$ be an arbitrary set of $n$ numbers, $S_1$, with normalization function $f_{S_1}$. Consider the set $S_2 = \{\frac{1}{a_i}\}$ and the set $S_3 = \{1_i\}$, which consists of $n$ copies of 1. Then condition (1) implies that, for all $i$, $f(c(f(c(\frac{1}{a_i}, a_i)), 1)) = f(c(\frac{1}{a_i}, f(c(a_i, 1))))$, where the various $f$'s are relative to the relevant sets. For example, in $f(c(\frac{1}{a_i}, a_i))$, $f$ is a rescaling function defined on the set $\{c(\frac{1}{a_i}, a_i)\}$ . Note that we are abusing notation here: strictly speaking the various $f$'s are not the same function, since they are defined over different sets. However, to avoid needless clutter, I use $f$ without subscripts.

According to the characterization of the combination function, the combination function is either multiplicative or additive. Since the derivations are very similar, I will only show that the normalization function must be multiplicative given that the combination function is multiplicative. So suppose that the combination function is $c(a, b) = ab$. Then we get: $f(f(\frac{1}{a_i} * a_i) * 1) = f(\frac{1}{a_i} * f(a_i * 1))$. Thus, we have: $f(f(1)) = f(\frac{1}{a_i} * f(a_i))$, i.e. $f(\frac{1}{a_i} * f(a_i))$ is a constant. But since, $f$ is one-to-one, that means $\frac{1}{a_i} * f(a_i)$ must also be a constant. That is, there exists a constant $k$ such that, for all $a_i$ in $S$, $\frac{1}{a_i} * f(a_i) = k$. Hence $f(a_i) = k * a_i$ for all $a_i$. Since $S$ was an arbitrary set, it follows that in general the normalization procedure must be multiplicative given that the combination function is multiplicative.

# C Characterization of inferential updating

The goal in this section is to show that the only legitimate updating rule that satisfies Regularity is inferential updating. According to the results in sections 3.1 and 3.2, any legitimate updating rule must either have (1) a multiplicative combination step and 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 $h_i$:

$$N(e_i + h_i) > 0 \tag{C.1}$$

Since the normalization function is assumed to satisfy (2), C.1 implies that the following is true for all $i$, where $d$ is an additive normalization constant:

$$e_i + h_i + d > 0 \tag{C.2}$$

Since the posterior probabilities must sum to 1, we also have:

$$\sum_i (e_i + h_i + d) = 1 \tag{C.3}$$

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

$$e_i + h_i - \frac{1}{n}\sum e_i > 0 \tag{C.4}$$

But it's obvious that (D.4) will not in general be true. For example, suppose $e_1$ is the smallest $e_i$. Then $r = e_i - \frac{1}{n}\sum e_i < 0$. Now suppose it's also the case that $h_1 < -r$. Then we have:

$$e_1 + h_1 - \frac{1}{n}\sum e_i = r + h_1 < 0 \tag{C.5}$$

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

# D 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:

$$p_E(H) = \begin{cases} 0 & \text{Given that } x \text{ is sufficiently low} \\ h_i + e_i + d & \text{Otherwise} \end{cases} \tag{D.1}$$

If the updating rule is conservative, then as few hypotheses as possible should be assigned a posterior probability of 0. 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 0. In order for the posterior probabilities to be probabilistic, we must have:

$$\sum_{i>m}(h_i + e_i) + (n - m)d = 1 \tag{D.2}$$

Now suppose we have a different updating rule resulting in some posterior $p'$ that is *not* conservative: i.e. there 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 $d'$:

$$\sum_{i>m'}(h_i + e_i) + (n - m')d' = 1 \tag{D.3}$$

Comparing D.2 and D.3 and remembering that $m' > m$, we see that:

$$0 < \sum_{i=m}^{m'}(h_i + e_i) = (n - m')d' - (n - m)d \tag{D.4}$$

And hence,

$$d < \frac{n - m'}{n - m}d' < d' \tag{D.5}$$

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 D.4 again. If both updating rules are conservative, then we have $m = m'$, and hence – making the necessary amendments in D.4, we have:

$$0 = \sum_{i=m}^{m'}(h_i + e_i) = (n - m)d' - (n - m)d \tag{D.6}$$

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.

# E 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:

$$p(H|E_1, E_2) = k * \text{Ev}[E_1|H, E_2]\text{Ev}[E_2|H]p(H) = k * f(L(E_1, H))f(L(E_2, H))p(H) \tag{E.1}$$

But we also have:

$$p(H|E_1, E_2) = k * \text{Ev}[E_1, E_2|H]p(H) = k * f(L(E_1, H) + L(E_2, H))p(H) \tag{E.2}$$

Comparing C.1 and C.2, 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.

# F 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 favors 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 favoring 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 $\bullet$ to denote the combination function, i.e. $e_1 \bullet e_2 = c(e_1, e_2)$, then it is plausible that $(G, \bullet)$ satisfies the following axioms:

1. **Closure.** For all possible evidential scores $e_1$ and $e_2$, $e_1 \bullet e_2$ is also a possible evidential score.

2. **Associativity.** For all possible evidential scores $e_1$, $e_2$ and $e_3$, $(e_1 \bullet e_2) \bullet e_3 = e_1 \bullet (e_2 \bullet e_3)$.

3. **Identity.** There exists a possible evidential score $i$ such that for all $e$, $i \bullet e = e \bullet i = e$. I.e., there exists a real number that represents evidence that has no effect (either favorable or unfavorable).

4. **Inverse.** For each possible evidential score $e$, there exists a possible evidential score $e'$ such that $e \bullet e' = e' \bullet e = i$. I.e. every evidential score could potentially be cancelled out by other countervailing evidence.

5. **Commutativity.** For all possible evidential scores $e_1$ and $e_2$, $e_1 \bullet e_2 = e_2 \bullet e_1$. I.e. the order in which the evidence is considered is irrelevant.

6. **Archimedean property.** For all possible evidential scores $e_1$ and $e_2$, there exists an integer $n$ such that $e_1 < e_2 \bullet e_2 \ldots \bullet e_2(\mathrm{n}times)$.

It is also plausible that the set of evidential scores is totally ordered: for all evidential scores $e_1$ and $e_2$, either $e_1 > e_2$ or $e_1 \leq e_2$. If we assume that the set of evidential scores form a totally ordered Archimedean group, then we can use the

following important result from group theory (see (Kopytov and Medvedev, 1996, p. 33), 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*:
>
> 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*:

1. If the normalization function is additive, then $c(x, y) = x + y$.
2. 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 $e^{-1}$:

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

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, (F.1) 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

$$a = \frac{1}{\sum e_i} \tag{F.2}$$

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 $(0, \frac{1}{ae})$, i.e. 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:

$$g(yx) = g(x) + K \tag{F.3}$$

Where $K$ is a constant that depends on $y$, but not on $x$. Interchanging the role of $y$ and $x$, we also have:

$$g(xy) = g(y) + K' \tag{F.4}$$

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

$$g(xy) = g(x) + g(y) + C \tag{F.5}$$

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, (F.5) implies that $g(i) = g(1 * i) = g(1) + g(i) + C$. Thus $g(1) = -C$. Using (F.5) 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 $\frac{1}{i} = 1$, i.e. $i = 1$. Hence $-C = g(1) = g(i) = 0$, so $C = 0$. Finally, then, we have, for all $x > 0$ and $y > 0$:

$$g(xy) = g(x) + g(y) \tag{F.6}$$

Now put $r(x) = g(e^x)$. Then (F.6) becomes, for all real $x$ and $y$:

$$r(x + y) = r(x) + r(y) \tag{F.7}$$

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:

$$x \bullet y = f(g(x) + g(y)) = e^{(\log(x^c) + \log(y^c))^{\frac{1}{c}}} = e^{(c * \log(xy))^{\frac{1}{c}}} = xy \tag{F.8}$$

I.e. the combination function is multiplicative, $c(x, y) = xy$.