On the Ecological and Internal Rationality of Bayesian Conditionalization and Other Belief Updating Strategies

Olav Benjamin Vassend

Abstract

According to Bayesians, agents should respond to evidence by conditionalizing their prior degrees of belief on what they learn. The main aim of this paper is to demonstrate that there are common scenarios in which Bayesian conditionalization is less rational—both from an ecological and an internal perspective—than other theoretically well-motivated belief updating strategies, even in very simple situations and even for an "ideal" agent who is computationally unbounded. The examples also serve to demarcate the narrow conditions under which Bayesian conditionalization is guaranteed to be ecologically optimal. A second aim of the paper is to argue for a broader notion of rationality than what is typically assumed in formal epistemology. On this broader understanding of rationality, classical decision theoretic principles such as expected utility maximization play a less important role.

1 Introduction

Although there are notable alternatives,[1] the most influential formal framework in epistemology is Bayesianism. According to Bayesians, a rational agent's degrees of belief should be summarizable in the form of a probability function, and whenever new evidence is learned, the agent should "conditionalize" and set the new posterior probability of any given proposition to equal the prior conditional probability of the proposition given the evidence.[2] Conditionalization is just one of a number of possible strategies that agents can follow when they are deciding how to respond to evidence, but Bayesians claim that

[1] For example, ranking theory (Spohn, 2012) and the AGM model of belief revision (Alchourrón et al., 1985).

[2] A more formal statement of conditionalization is given in Section 2.

it is the only strategy that is rational—at least for ideal agents who are unencumbered by computational or time constraints. The main goal of this paper is to refute the preceding claim and to formulate alternative well-justified updating rules that are more rational than Bayesian conditionalization under plausible—indeed common—conditions. At the same time, the paper aims to argue for an account of rationality that is broader and more context-dependent than the account typically assumed among formal epistemologists. On the broader account, classical decision theoretic principles like dominance and expected utility maximization will be shown to play a more limited role.

Loosely following Gigerenzer and Todd (2012) and Todd and Brighton (2015), I will say that a strategy is "ecologically rational" (epistemically or practically) for an agent in proportion to how effective the strategy is in helping the agent attain its goals in the given environment. Note that this characterization conceives of rationality as something that comes in degrees rather than being a binary property. Note also that the characterization is somewhat vaguely stated, since what it means for a strategy to be "effective", in particular, is left imprecise. In general, the appropriate precisification is context-dependent.

Ecological rationality is an "externalist" notion of rationality in the sense that whether an agent is rational is determined by whether the agent's behavior in fact is effective in the agent's environment, regardless of what the agent may believe. As a complementary notion, I will therefore say that a strategy is "internally" rational for an agent if and only if the strategy appears to be ecologically rational from the agent's point of view.[3] As in the case of ecological rationality, I will argue later that the way in which internal rationality should be spelled out is context-dependent.

The preceding goals-oriented and context-sensitive characterizations of rationality may be contrasted with a different conception that is popular in epistemology. Many proponents of Bayesianism hold that updating one's degrees of belief in a way that violates Bayesian conditionalization is irrational in a manner that is similar to the way in which believing a proposition and its negation at the same time is irrational. That is, it is irrational, not primarily because it will lower your chance of attaining your goals (although most Bayesians think it will), but because it is a fundamentally incoherent way of responding to evidence. On this view, Bayesian conditionalization is not merely an instrumentally valuable tool that happens to work well in a wide variety of circumstances; instead, it is a structural norm that is binding even if following the norm would lead you to be less successful. Furthermore, this conception of rationality is binary: an agent whose degrees of belief violate the probability axioms is not merely less rational than an agent whose degrees of belief do not; instead, such an agent is irrational, full stop.[4] I will call this context-independent and

[3]Easwaran (2021) also draws a distinction between ecological and internal rationality. Easwaran's distinction roughly maps onto the distinction I draw below between ecological and formal rationality.

[4]Although there are notable attempts to make this kind of conception of rationality graded rather than

binary conception of rationality "formal rationality."

Bayesians are well aware that formal and ecological rationality can come apart (Greaves, 2013). For example, suppose I give you a gold coin every time you have degrees of beliefs that are non-probabilistic. Presuming your (only) goal is to amass as many gold coins as possible, the ecologically rational strategy for you will probably be to have degrees of belief that are non-probabilistic, even though doing so may be formally irrational.

There are other, more ordinary examples where ecological and formal rationality point in different directions. In particular, Bayesian computations are very time-consuming and difficult. Hence, in practice it may often be ecologically rational to use heuristics rather than rigorous calculations to form informed degrees of belief and make decisions (Tversky and Kahneman, 1974, Marsh et al., 2004, Vranas, 2000, Gigerenzer and Todd, 2012).[5] Along similar lines, Bacchus et al. (1990) give examples that show how learning a piece of evidence may lead (bounded) agents to realize that their earlier expectations of what evidence they would see were mistaken, which may in turn make it rational for those agents to change their beliefs in a way that disagrees with Bayesian conditionalization. Because of counter-examples of this sort, it is not uncommon to maintain that Bayesian norms are optimal for "ideal" reasoners who are not encumbered by computational limitations, while conceding that more limited agents may sometimes need to take shortcuts or otherwise deviate from Bayesian conditionalization. However, insofar as these shortcuts and deviations are justified, Bayesians often maintain that this is in part because they approximate the Bayesian ideal.

By contrast, the main goal of this paper is to show that Bayesian updating is sometimes not ecologically optimal even for "ideal" reasoners. More particularly, the paper will argue that there are several commonplace scenarios, where computational limits are not an obstacle, in which there are updating rules that are more ecologically rational than conditionalization, and not just because they are imperfect approximations of a Bayesian ideal. The arguments in this paper are congenial to Douven (2020), who gives an example that he claims demonstrates that explanationist updating strikes a better balance between accuracy and speed, given an imagined scenario where the goal is to save patients in an Intensive Care Unit. Douven points out that, in his example, even an ideal agent (not computationally limited) would be better off not using Bayesian updating. However, Douven's example still presupposes that time is a resource constraint. By contrast, the examples presented in this paper will show that there are circumstances where Bayesian conditionalization is suboptimal even if there are no resource constraints of any kind.

---

binary; see, e.g., Zynda (1996), Bona and Staffel (2017), De Bona and Staffel (2018).

[5]Modern Bayesian statistics is itself only possible because of sophisticated approximations and quite recent advances in computing. But even with modern computing power, Bayesian calculations are often too demanding.

In order for the examples to be convincing, they arguably need to satisfy two requirements. First, the way we "score" the extent to which a strategy is rational should not be rigged against Bayesian conditionalization in the way that the above-mentioned example (where I pay you to have non-probabilistic degrees of belief) is rigged against probabilism. Indeed, Pettigrew (2021) argues that some of Douven's examples are flawed in precisely this way. Second, any alternative updating rule that is purported to be ecologically more rational than Bayesian conditionalization should have some independent theoretical motivation. It should not just be an ad hoc updating rule that has been designed to beat Bayesian conditionalization in one particular situation; instead, there should be clear, principled reasons why we would expect the updating rule to be more ecologically rational than Bayesian conditionalization in a wide range of similar scenarios, and the proposed updating rule should ideally be grounded in a sound theoretical framework. This paper will use a "minimum divergence framework" that has been gaining prominence in several disciplines in recent years as a general metaframework for motivating and comparing updating rules.

It is worth noting that there is growing empirical evidence from statistics that Bayesian conditionalization can perform poorly in applied contexts. For example, if data are "overdispersed" —i.e., the variance exhibited by the data is greater than that expected by one's statistical model—then Bayesian conditionalization can result in posterior distributions that are overly confident (Holmes and Walker, 2017).[6] Grünwald and van Ommen (2017) give examples from linear regression that show that Bayesian conditionalization can yield poor predictions. Finally, using Bayesian conditionalization to assign a probabilities to statistical models has serious challenges, as the resulting posterior probabilities are often very sensitive to the prior probabilities assigned to parameters inside the models. Yao et al. (2018) show that a technique they call "Bayesian stacking" (and which will be discussed below) often produces dramatically better predictions than standard Bayesian conditionalization.

Epistemologists may be apt to dismiss the empirical shortcomings of Bayesianism as practical issues in applied statistics that have little relevance to the standing of Bayesianism as a normative framework for epistemology. By analogy, utilitarianism is hard to apply in many of the kinds of complex moral quandaries that arise in practice, but that does not mean that utilitarianism is not ultimately the correct theoretical framework for ethical reasoning.[7] I personally think this kind of attitude is a mistake (both in epistemology and ethcis), but I concede that it has some force. Hence, my goal is to provide simple—but representative—examples where Bayesian conditionalization is less ecologically and

[6] "Overly confident" in the sense that the posterior distribution is much more sharply concentrated than it would be if one had the true statistical model.

[7] I am grateful to Frank Cabrera for emphasizing this analogy to me.

internally rational than other well-motivated belief updating strategies.

The plan of the paper is as follows. Section 2 introduces the minimum divergence framework. Section 3 argues that Bayesian conditionalization is not always ecologically optimal or internally rational if we are unsure whether the hypotheses under consideration form a partition. Section 4 gives a simple example that shows that Bayesian conditionalization can be ecologically suboptimal and internally irrational even if we know that the hypotheses under consideration form a partition. Section 5 details the conditions under which Bayesian conditionalization is guaranteed to be ecologically optimal. Finally, Section 6 ends the paper with a few concluding remarks.

2 A minimum divergence perspective on Bayesian updating

Bayesian conditionalization says that an agent's posterior degrees of belief should be related to the agent's prior conditional degrees of belief in the following way:

> *Bayesian conditionalization:* If p is a probability function that quantifies a rational agent's degrees of belief, $p(H)$ is the agent's prior degree of belief in H, and the agent learns E (and nothing else), then the agent's posterior degree of belief, $p_E(H)$, is equal to the agent's prior degree of belief in H conditional on E:
>
> $$p_E(H) = p(H|E) \tag{2.1}$$

In practice, the agent's prior conditional degree of belief in H given E, $p(H|E)$, is usually calculated via the following version of Bayes's formula, which relates $p(H|E)$ to the likelihood $p(E|H)$ and prior probability $p(H)$ of H:

$$p(H|E) = \frac{p(E|H)p(H)}{\sum_i p(E|H_i)p(H_i)} \tag{2.2}$$

Note that the denominator of 2.2 assumes that H belongs to a partition, i.e., an exhaustive set of mutually exclusive propositions $\{H_i\}$.

Although there are several justifications for Bayesianism, a particularly powerful and general foundation is provided by the minimum divergence framework ((Bernardo (1979), van Fraassen (1981), Diaconis and Zabell (1982), Berger et al. (2009), Bissiri et al. (2016), Eva and Hartmann (2018)). In the minimum divergence framework, belief updating is regarded as a balancing act: when we receive new evidence, we should adjust our beliefs,

but we should do so gradually and incrementally; we should not let our beliefs be completely dictated by whatever the latest news is. In Bayesian terminology, what we want is for our posterior degrees to belief to stay relatively close to our prior degrees of belief, all while according due weight to our evidence. The minimum divergence framework formalizes the preceding ideas mathematically. The starting point is a classical result that traces back to Williams (1980) and was generalized by Diaconis and Zabell (1982), which shows that Bayesian conditionalization uniquely solves the following optimization problem: find the joint posterior distribution that is as close as possible—in terms of a certain measure of statistical distance called the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951)—to the joint prior distribution, subject to the constraint that the posterior probability of the evidence equal one. More formally, the optimization problem can be formulated as finding the posterior probability function p_E that minimizes the following expression, subject to the constraint that $p_E(E) = 1$:

$$KL(p_E, p) = \sum_i \sum_j p_E(H_i, E_j) \log \frac{p_E(H_i, E_j)}{p(H_i, E_j)} \tag{2.3}$$

If we use the fact that $p_E(E) = 1$, we can simplify the above expression and write it in the following alternative manner (indeed, some authors (e.g., Bissiri et al. (2016)) have 2.4 as their starting point):

$$Optim(p_E) = \sum_i p_E(H_i) \log \frac{p_E(H_i)}{p(H_i)} - \sum_i p_E(H_i) \log p(E|H_i) \tag{2.4}$$

Here, we see that the KL divergence between the joint posterior and joint prior is really made up of two components: the first component is the KL divergence between the posterior and prior probability distributions over the hypotheses under consideration, while the second component is a sum (weighted by the posterior distribution) of the log likelihood of each hypothesis on the evidence. Although 2.4 may look complicated—and perhaps no more illuminating than 2.3—it provides a valuable perspective on what the posterior distribution actually is. In words, it is a compromise between two different goals: the first goal is to stay close (in terms of Kullback-Leibler divergence) to the prior distribution—this is what the first term in 2.4 ensures; the second goal—quantified by the second term in 2.4—is to assign a high posterior probability to hypothesis that have accurately predicted the evidence, where predictive accuracy is measured by way of the "logarithmic scoring rule", which scores the predictive accuracy of a hypothesis, H on evidence E as $\log p(E|H)$. The Bayesian posterior distribution is the distribution that uniquely balances these two goals in an optimal way, given that the goals are quantified in terms of minimizing expression 2.4. Regarding the Bayesian posterior distribution as

a solution to a certain optimization problem is useful because it immediately suggests the possibility that different situations and different values may make us want to solve optimization problems other than 2.4. Below we will see concrete examples where this is arguably the case.

3 Case 1: The agent is unsure whether the propositions under consideration form a partition

As was mentioned earlier, calculating the quantity $p(H|E)$ usually involves using formula 2.2, which assumes that the hypotheses under consideration form a partition. However, in many situations—both everyday and scientific—it is not easy to determine whether this assumption is satisfied. For example, there are various purported explanations for why the dinosaurs went extinct, including that there was an asteroid strike and that the climate changed. A scientist who wanted to do a Bayesian analysis to determine which hypothesis is most plausible would have to assume that these explanations form a partition, but it is hardly unthinkable that there might be some other explanation that the scientific literature has not considered. In general, there is no way of knowing whether one of the hypotheses under consideration is true, unless the hypotheses form a logical partition that guarantees that one of them must be true (e.g., the hypotheses are of the form $\{A, \neg A\}$).

In the Bayesian statistical literature, the situation where we are not sure whether our hypotheses form a partition is called the "M-open" case (Bernardo and Smith, 1994); in the philosophical literature, it is sometimes known as the problem of unconceived alternatives (Stanford, 2006). In both literatures, it is recognized as a significant problem for Bayesianism,[8] both theoretically and in practice. Arguably, it is a problem that would arise even for a hypothetical computationally unbounded agent, unless the agent were literally omniscient—if the set of hypotheses do not form a logical partition, then it seems there is no way of knowing whether one of the hypotheses is true without knowing what the true hypothesis is.

There are two standard proposals for how Bayesians might solve the problem of unconceived alternatives. The first proposal is to add a a "catchall" alternative hypothesis to the set of hypotheses that says that none of them are true, i.e., $C = \neg(H_1 \vee H_2 \vee \ldots \vee H_n)$ Shimony (1970). Adding C to the set of hypotheses will logically guarantee that the set forms a partition, so Bayesian conditionalization can now proceed unhindered—at least in theory. However, there are major problems with this proposal: including C in the set of hypotheses means we will have to assign it a prior probability distribution and a likeli-

[8] Although it is by no means only a problem for Bayesianism.

hood on the evidence. It is by no means clear how this is to be done in a way that is not completely arbitrary (see, e.g., Chapter 1 of Sober (2008) for detailed discussion).

A second—and more promising—solution to the problem of the unconceived alternative is simply to concede that our degrees of belief are always relative to the set of hypotheses under consideration—our degrees of belief are usually conditional rather than unconditional (Salmon, 1990). Or, to use Sprenger's (2019) terminology, our degrees of belief are typically (implicitly if not always explicitly) suppositional rather than actual: thus, $p(H_i)$ is our degree of belief that H_i is true, conditional on the supposition that one of the hypotheses we are considering is true. Sometimes we are fairly confident that this supposition is true and sometimes we have no idea, but (outside of simple scenarios) we are rarely 100 $\%$ sure.

I think this response to the problem of unconceived alternatives is correct, but the following question now arises: should the way we manage our degrees of belief be sensitive to the possibility that none of the hypotheses under consideration is correct? Or, put differently, should suppositional degrees of belief be updated in the same way as actual degrees of belief? The standard (albeit usually implicit) answer among epistemologists appears to be "yes"—as far as updating one's degrees of belief is concerned, there is no distinction between knowing that something is true and supposing that it is true. More formally, we can phrase what is arguably the standard view as follows: If K represents the *knowledge* that our background assumptions are correct and S represents the *supposition* that our assumptions are correct, then: $p_E(H_i|S) = p_E(H_i|K)$.

This view is arguably standard among Bayesian statisticians as well, most of whom use Bayesian conditionalization in the M-open case. However, at the same time, Bayesian statisticians generally acknowledge that, if the posterior probability of a hypothesis is conditional on suppositions that may be wrong, then those suppositions—and inferences made on the basis of the supposition—need to be independently checked, since an inference made on the basis of a seriously mistaken assumption may be misleading. Thus, even if we grant that $p_E(H_i|S)$ and $p_E(H_i|K)$ should numerically agree, that does not mean we must hold that they have the same epistemic status: it is rational to maintain some degree of higher-level skepticism towards the former degree of belief. On this picture, Bayesian conditionalization plays a part in how one should manage one's degrees of belief in the face of the possibility that none of the hypotheses under consideration is true (or even close to true), but it is only a part of the story: Bayesian conditionalization must be supplemented with independent (non-Bayesian) checks.[9]

An alternative (and complementary) idea is that suppositional and actual degrees of

[9]This perspective is adopted in influential textbooks in Bayesian statistics, including Gelman et al. (2013) and McElreath (2016). An accessible and brief overview of this approach to Bayesian statistics is given in Gelman and Shalizi (2013).

belief should not necessarily be updated in the same way. The idea that it might sometimes be rational to update probabilities in a non-Bayesian way has been gaining traction in Bayesian statistics in recent years (e.g., Zhang (2006), Bissiri et al. (2016), Grünwald and van Ommen (2017)), and there are several papers that appear to show that Bayesian updating may be improved upon in certain contexts and given certain inferential or predictive goals agents may have. Bissiri et al. (2016) attempt to give a general framework for how to update probability distributions in a way that replaces the likelihood with a user-specified loss function. In the philosophical literature, Douven (2013, 2016) and Douven and Wenmackers (2017) argue that an updating method based on Inference to the Best Explanation is often better than Bayesian conditionalization, given certain aims agents may have. Vassend (2019) attempts to give a general framework that subsumes the framework in Bissiri et al. (2016) as well as other non-Bayesian updating rules, such as updating based on Inference to the Best Explanation.

The rest of this section is in the spirit of this recent work. The goal is to show a concrete example of a theoretically well-motivated updating rule that we have strong reasons to believe will be more rational than Bayesian conditionalization in certain simple and mundane situations in which we are unsure whether our hypotheses form a partition. In the type of scenario that will be discussed, the focus is not primarily on finding which hypothesis H_i is true; instead, the goal is to make accurate predictions about various states of the world, and the hypotheses under consideration are regarded as tools for prediction rather than being of intrinsic interest. This is not some sort of esoteric scenario. Much of our thinking both in everyday life and—especially—in the sciences proceeds in this way: we are wondering whether some state of the world S is true (e.g., whether it will rain tomorrow, whether the butler is guilty, whether someone is lying to us), but we have no way of assessing S's plausibility directly. Hence, we form alternative hypotheses, each of which confers a (more readily assessable) probability on S, and then we evaluate the relative plausibility of the competing hypotheses in response to evidence and form a final assessment of S's plausibility by using the law of total probability, i.e., $\sum_i p(S|H_i)p_E(H_i)$.

The preceding points may be obvious, but I emphasize them because some Bayesians may be inclined to claim that if it is S that is of interest, then we should just condition $p(S)$ on E directly and not go via a set of hypotheses H_i. In practice, however, our reasoning will often necessarily involve hypotheses that are not, in themselves, of interest, but which we must use in order to form informed degrees of belief about the propositions we care about. In fact, there are good reasons for thinking that prediction is necessarily mediated by theories and models. Sterkenburg and Grünwald (2021) argue that few (perhaps none) of the standard machine learning algorithms are "purely data-driven," in the sense of providing a prediction of states of the world S on the basis of evidence E alone. Instead, they generally require a model of some sort as an additional input. Furthermore,

this is not a coincidental state of affairs, but is instead due to the fact that no prediction method works well under all circumstances (this is known as the "no-free-lunch" theorem (Wolpert, 1996)). Hence, there are good reasons for thinking that even an ideal reasoner—logically omniscient and computationally unbounded—would still reason and form predictions in a model-mediated manner.

Let me be clear in saying the claim is not that Bayesian conditionalization is always ecologically irrational if it is uncertain whether the true hypothesis is under consideration—I believe there are probably many scenarios where Bayesian conditionalization is ecologically rational, even if the standard assumptions of Bayesian conditionalization are not satisfied. The claim is just that there are certain cases where Bayesian conditionalization will not be ecologically optimal, for reasons that are relatively easy to understand. And because those cases can be anticipated by agents, Bayesian conditionalization is also internally suboptimal in these examples.

To start, note that an important property of Bayesian conditionalization is that–as the amount of evidence increases—it will almost always concentrate all of the posterior probability on the single hypothesis that has been most predictively accurate. This is clear from 2.4 because as the evidence accumulates, 2.4 will be increasingly dominated by the second term, and the second term—in turn—will be increasingly dominated by the single hypothesis that has the best log score on the evidence. Thus, in the limit, Bayesian conditionalization converges on the single hypothesis that has the best log score. But what we are interested in is finding the posterior distribution that will result in the best possible *predictive* distribution, i.e., we would like $\sum_i p(S|H_i)p_E(H_i)$ to be maximally accurate (e.g., have an optimal log score). If one of the hypotheses under consideration is true, then these two goals coincide: the best predictive distribution is simply the one that assigns all its probability to the truth.[10] But if the truth is not under consideration or we are not sure whether it is, then the two goals may diverge.

For example, suppose that we have two rival weather models, m_1 and m_2, that we will use to predict the probability of rain, R. The Bayesian prediction is $p(R|m_1, E)p(m_1|E) + p(R|m_2, E)p(m_2|E)$ and requires us to come up with posterior (suppositional) degrees of beliefs in m_1 and m_2. Suppose it turns out that, on the evidence so far, one of the models has systematically overpredicted the probability of rain while the other model systematically has underpredicted the probability of rain. If there is a lot of evidence, Bayesian conditionalization will still concentrate all its probability on one of the hypotheses—whichever has the best log score on the evidence—and will consequently dictate that our (suppositional) degree of belief in that hypothesis should be 1 (or close to 1). But this does

[10] As long as we assume that the hypotheses are fully specified—i.e., they are not statistical models that contain adjustable parameters. On limited data, a true model with adjustable parameters may well not be the most predictively accurate model available (Forster and Sober, 1994).

not seem reasonable. Given that our evidence so far strongly suggests that the two models are biased in opposite directions, it seems better to strike some sort of balance: we should have a higher degree of belief in the model that has been more accurate, but we should still maintain some degree of belief in the less predictively accurate model, and our best estimate of whether it is going to rain in the future should be some compromise between the predictions made by the two models.

Looking at 2.4 again makes it clear that Bayesian updating does not actually solve the optimization problem in which we are primarily interested, at least not directly, if what we would like to do is maximize the accuracy of the posterior predictive distribution. Formally, the optimization problem we really want to solve is to find the posterior distribution that minimizes the following expression:

$$Optim(p_E) = \sum_i p_E(H_i) \log \frac{p_E(H_i)}{p(H_i)} - \log \sum_i p_E(H_i) p(E|H_i) \tag{3.1}$$

The difference between 2.4 and 3.1 is subtle, but important: in the second term of 2.4 we are scoring the posterior distribution p_E by its ability to assign high probabilities to *hypotheses* that have a good log score on the evidence; in 3.1, on the other hand, we are scoring p_E by its ability to produce a *predictive distribution* that has a good log score on the evidence. It is also possible to consider generalizations of 3.1, where, for example, we assign different weights to the first and second terms. The problem of finding the posterior that minimizes 3.1 is closely related to what Yao et al. (2018) refer to as the Bayesian "stacking problem": the Bayesian stacking problem, as discussed in Yao et al. (2018), is the problem of finding the probability distribution that minimizes the *second* term in 3.1. Because the stacking problem may be regarded as a special case of 3.1 (where we are assigning the first term of 3.1 a weight of 0), I will refer to the posterior distribution that minimizes 3.1 as the "stacking posterior." In Yao et al. (2018), Bayesian stacking is presented as an ad hoc way of averaging multiple predictive distributions that is empirically superior to Bayesian conditionalization. A major point in favor of the minimum divergence framework is that it provides a theoretical unification of the two methods, since we can see that Bayesian conditionalization and Bayesian stacking may in fact be regarded as solving closely related (but still importantly different) optimization problems.

If the true hypothesis is not under consideration, then the stacking posterior may well be different from the Bayesian posterior. On the other hand, if the true hypothesis is under consideration, then—as more and more evidence accumulates—the posterior that concentrates all of its probability on the true hypothesis will optimize both 2.4 and 3.1. Overall, then, we have reason to think that optimizing 3.1 is the more *robust* option if we are unsure whether the true hypothesis is under consideration in the sense that the stacking posterior predictive distribution will be roughly as accurate as the Bayesian posterior predictive dis-

tribution if the true hypothesis is under consideration and that it will be more accurate otherwise.

Let us look at a simple example, where the hypotheses under considerations concern the chance that some event will occur.[11] To make the example more concrete, we can think of the hypotheses as being about the possible biases of a coin. We consider the following four hypotheses about the value of the bias: $\{0.05, 0.35, 0.65, 0.95\}$. We now consider two different ways of generating the true value of the bias. In the first, *well-specified* case, the true bias is selected (with uniform probability) to be one of the four hypotheses under consideration. In the second, *misspecified* case, the true bias is selected, with uniform probability, to be any number between 0 and 1. Note that in the misspecified case there is a probability of 1 that the true bias of the coin will not be one of the hypotheses under consideration. Given the true value of the bias, we generate n training data points and 1000 test data points, where n ranges from 1 to 200. We assume that the prior over the four hypotheses under consideration is uniform and we update the prior on the evidence using Bayesian conditionalization and stacking updating. Figure 1 compares the log score of the Bayesian posterior predictive distribution and the stacking posterior predictive distribution as a function of n in both the well-specified and misspecified situations. Note that each data point in the figure is a result of averaging the results from 100 independent executions of the simulation.

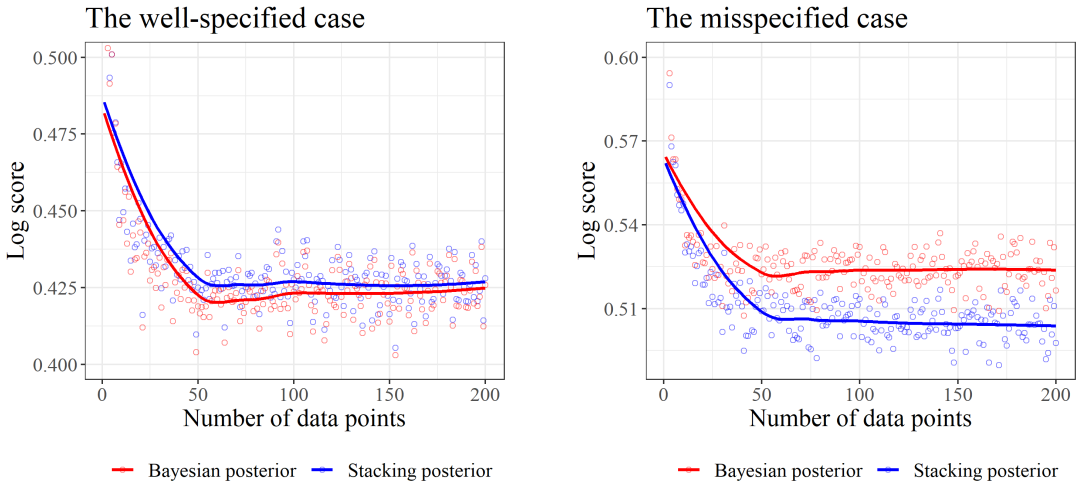

Figure 1: The Bayesian posterior predictive distribution vs the stacking posterior predictive distribution in two scenarios.

[11] All the simulations in this paper are done using the statistical programming language R (R Core Team, 2020).

In the well-specified case, both the Bayesian posterior and the stacking posterior converge on having the same accuracy, although the Bayesian posterior does so somewhat faster. In the misspecified case, the stacking posterior and the Bayesian posterior do not converge on having the same accuracy, and the stacking posterior is clearly superior. These results are in line with what we would expect based on the preceding theoretical discussion. For an agent who is in the misspecified case, the stacking posterior is straightforwardly more ecologically rational than the Bayesian posterior, since it yields more accurate predictions on average for every possible sample size. It is also clear that, from the point of view of an agent who knows that they are in the misspecified case, the stacking posterior will be more rational than the Bayesian posterior regardless of the agent's prior distribution, provided that the agent knows that they will receive a large number of data. This is because asymptotically—regardless of the agent's prior distribution over the hypotheses under consideration—the results in Figure 1 will be the same: i.e., in the misspecified case, the stacking posterior will converge on a predictive distribution that is more predictively accurate than the Bayesian predictive distribution. Finally, if the agent does not know whether they are in the misspecified or well-specified case—i.e., they are in the "M-open" case—the stacking posterior again seems more internally rational than the Bayesian posterior, because the stacking posterior is only slightly less accurate in the well-specified case, but much more accurate in the misspecified case.

The above conclusions might seem inconsistent with various arguments that purport to show that Bayesian updating is uniquely rational. For reasons of space, I cannot discuss each of these arguments in detail, but for the purposes of illustration, let me single out one of the most prominent ones. If we assume that expected utility maximization is rational, then a result due to Greaves and Wallace (2006) purports to show that, for any agent, Bayesian conditionalization is uniquely rational from that agent's point of view because it has a higher expected utility than any other updating rule, provided that the expectation is calculated relative to the agent's own prior distribution.[12] However, closer inspection shows that Greaves and Wallace's (2006) argument assumes that the hypotheses under consideration (or states of the world) form a partition, and this assumption is violated in the misspecified case. Indeed, any expectation over a set of possibilities implicitly assumes that one of the possibilities is true, and if this assumption is violated the relevance of the calculation is clearly in doubt. Suppose I will roll a die with faces numbered 1 through n several times and I ask you to estimate the average number it will land on. On the assumption that it is a normal die with six faces, the expected value of the die roll is 3.5, and this is a good guess as to the average number the die will show in the long run. But if I tell you that the die does not have six faces, the calculated expectation is not necessarily

[12] And provided the utility function is "proper"—this condition is discussed in the next section of this paper.

a good basis on which to formulate predictions or make decisions. It might not be useless, provided that the actual number of faces is close to six. But it is clearly not rational to wager a significant amount of money on your guess, because your guess is based on an assumption you know to be wrong.

The more general point is that expected utility calculations are not necessarily a good guide for rational decision making and belief formation if the calculations are based on incorrect assumptions, and this same point carries over to the case where you do not *know* whether the requisite assumptions are satisfied, e.g., if you are in the "M-open case" and do not know whether the possibilities under consideration form a partition. In either case, expected utility calculations should be taken with a grain of salt, and for that reason Greaves and Wallace's (2006) argument arguably loses much of its force.

Of course, Greaves and Wallace's (2006) argument is not the only one that has been advanced on behalf of conditionalization. A more recent argument shows that Bayesian conditionalization "dominates" all other updating rules, where a strategy is said to dominate another one if and only if it is better (or at least not worse) in all possible worlds according to some metric (Pettigrew, 2021). For reasons of space, I will not go into a detailed discussion of the argument. However, I will point out that dominance reasoning—like expected utility maximization—is arguably suspect if we do not know whether we are considering a partition of possibilities. The fact that A dominates B across a set of possibilities under consideration does not necessarily mean that A will be ecologically more rational than B, if we have neglected to consider important possibilities. An alternative to using either expected utility maximization or dominance reasoning—both of which are arguably not robust in either the misspecified or M-open case—is to use a more informal and context-specific method to evaluate and compare the merit of various strategies, as we did a few paragraphs ago.

The example in this section is admittedly simple. However, one might have expected Bayesian updating to have an especially good shot at being ecologically optimal precisely because the example is so simple: this is not a computationally complex example where a Bayesian calculation is intractable, nor is it hard to come up with well-justified priors or likelihoods. Indeed, the flat prior is arguably the objectively optimal prior over the given hypothesis space, because the four hypotheses do, indeed, have the same probability (density) of being chosen. In any case, the basic results will remain the same regardless of which prior distributions we use because the biggest difference between the Bayesian posterior and the stacking posterior in the misspecified case shows up when the amount of data gets large, i.e., when the influence of the prior becomes small.

As was mentioned earlier, 3.1 is a generalized version of what Yao et al. (2018) refer to as the Bayesian stacking problem. Yao et al. (2018) give several examples that show that Bayesian stacking is superior to Bayesian conditionalization, if the goal is to maxi-

mize predictive accuracy. Bayesian stacking, in turn, is an instance of a broader class of statistical model combination techniques referred to simply as "stacking," which traces its roots to Wolpert (1992) and Breiman (1996) and is a state-of-the art method for combining predictive models in statistics and machine learning. Given these facts, it is not surprising that the stacking posterior would outperform the Bayesian posterior in the simple example we have considered.

To summarize, if we are unsure whether one of the hypotheses under consideration is true, it may not be rational to concentrate all of our degree of belief on the single hypothesis that has proven to be best, which is what Bayesian conditionalization will often lead us to do. The stacking posterior is a principled way of updating suppositional degrees of belief in a way that is more cautious. Theoretically, the stacking posterior has a firm grounding in the minimum divergence framework; empirically, it outperforms Bayesian conditionalization in many scenarios, including the simple example provided in this section. Overall, then, there is strong reason to think that it is more rational than Bayesian conditionalization—both ecologically and from an internal perspective—in the kind of case we have discussed in this section.

4 Case 2: The agent values certain regions of the posterior distribution over others

Suppose we know that one of the propositions we are considering is true. Is Bayesian updating guaranteed to be ecologically optimal in this case? Not necessarily. To demonstrate how Bayesian conditionalization may fail to be ecologically optimal even under such favorable conditions, we can use the same example as in the previous section, except we now let the set of hypotheses consist of 100 equally spaced numbers from 0.01 to 0.99, one of which is the true bias, b. Suppose our goal is to maximize the accuracy of our estimate of the value of the bias, where we will use as our estimate the posterior expected value of the bias. In mathematical terms, then, our goal is to maximize the accuracy of $\mathrm{E}_{p_E}[H] = \sum_i H_i p_E(H_i)$, where the possible values of H_i are the possible values of the true bias. A natural way of measuring the accuracy of a proposed estimate is to take the absolute distance between the estimate and the actual value of the bias, i.e. $|\mathrm{E}_{p_E}[H] - b|$.[13] As in the previous section, we suppose that the prior distribution is uniform over the set of possible biases, which we now know form a partition. Given the results from the preceding section, we know that the Bayesian posterior is going to be better than the stacking

[13]There are many other measures of accuracy one might use, but my experience is that—for our current purposes—they all give the same qualitative results.

posterior in this kind of example, but is there another updating rule that could beat both? I will argue that the answer is yes, provided that we have certain values.

More precisely, suppose that we care more about the value of the bias if its value is close to either 0 or 1. Given these priorities, there is reason to think that Bayesian updating might not to be optimal. As 2.4 makes clear, Bayesian conditionalization implicitly penalizes hypotheses by their logarithmic score on the evidence, and the logarithmic score is very sensitive to small prediction mistakes. For example, if a hypothesis assigns an event that happens a probability of 0.1, then the hypothesis receives a log score of 2.3; if it assigns the event a probability of 0.01, then its score is instead 4.6—twice the penalty. The log score therefore has a built-in conservatism in that it will usually prefer hypotheses that do not assign very low or very high probabilities to events, unless there is a lot of evidence.[14] Not all scoring rules have this property. For example, if a probability distribution (over a binary partition) assigns a probability of f to the event that happens, then the quadratic score—also known as the Brier score—of the distribution is $(1-f)^2$. Note that if we use the Brier score, the penalties assigned to $f = 0.1$ and $f = 0.01$ are much closer in magnitude. There is therefore reason to think that the Brier score will be somewhat more efficient than the log score at finding the value of the true bias, if that value is close to 0 or 1. It is easy to verify in simulations that this is the case (and, indeed, the simulations I provide below can be taken to demonstrate precisely this fact). Let us therefore create an updating rule that is based on the quadratic score rather than the log score. We can do so easily by replacing the first term of 2.4 with the quadratic divergence between $p_E(H_i)$ and $p(H_i)$ and the second term of 2.4 with a weighted sum of the Brier scores of all the hypotheses on the evidence. That is, we consider the following optimization problem, where $B(H_i, E)$ is the Brier score of H_i on evidence E:

$$Optim(p_E) = \sum_i (p_E(H_i) - p(H_i))^2 - \sum_i p_E(H_i) B(H_i, E) \tag{4.1}$$

Let us call the posterior distribution that minimizes expression 4.1 the "quadratic posterior" and the associated updating rule "quadratic updating." Figure 2 plots the absolute distance between the true value of the bias and the quadratic posterior predictive estimate, and the absolute distance between the true value of the bias and the Bayesian posterior estimate, for every possible value of the true bias and given that the posterior distributions have been arrived at through updating the (uniform) prior distribution on 10 data points, 20 data points, 50 data points, and 100 data points. Note that each point in the figure is an average of 1000 executions of the simulation.

After 100 data points—or even just 50—the two posteriors converge on having roughly

[14] This feature of the log score is discussed in detail by Selten (1998) and Vassend (2018).

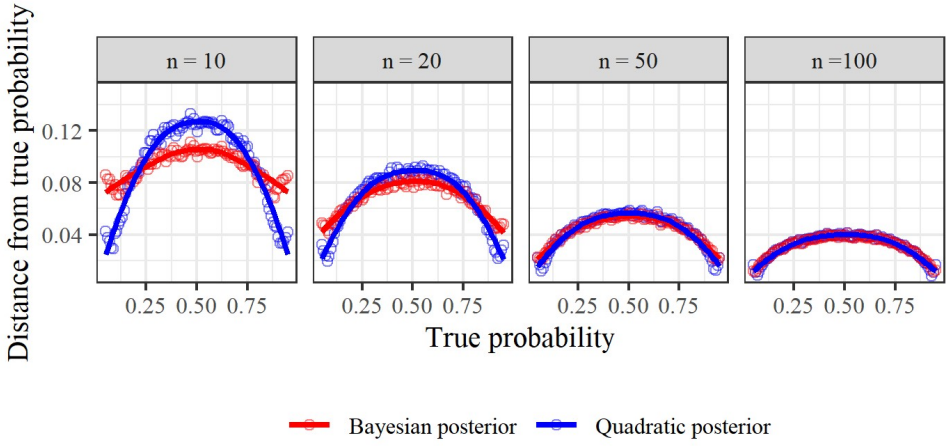

Figure 2: The distance of the Bayesian posterior predictive estimate and quadratic posterior predictive estimate from the actual probability, for each possible value of the actual probability.

the same accuracy, for every possible value of the bias. But for smaller data sets, it is clear that the two posteriors have a different profile, with the quadratic posterior being more accurate on the extremes (roughly for values of the bias below 0.2 or above 0.8) and the Bayesian posterior being more accurate in the middle of the range. If we value accuracy at the extremes of the range over accuracy in the middle of the range, we therefore have reason to prefer the quadratic posterior to the Bayesian posterior.

An objection might be that we should use whichever posterior predictive distribution has the highest expected accuracy, because in this example, we know that the hypotheses under consideration form a partition, and so the problems with using expected utility theory that were discussed in the preceding section do not apply. Since we are assuming that the prior is flat, the expected accuracy of each distribution can be calculated easily simply by averaging the accuracy of each distribution over all possible values of the true bias in Figure 2. It's straightforward to verify that Bayesian conditionalization has a higher expected accuracy than quadratic updating, if expected accuracy is calculated in this way. And, indeed, the aforementioned theoretical result due to Greaves and Wallace (2006) guarantees that this must be the case.

Nevertheless, I believe the fact that Bayesian conditionalization has a higher expected utility than quadratic updating is less impressive than it might seem for two important reasons. First, note that we are calculating the expected utility of each posterior with respect to our flat prior distribution. However, if the flat prior distribution does not reflect accurately the actual chances of events in the world, then a high expected utility with respect

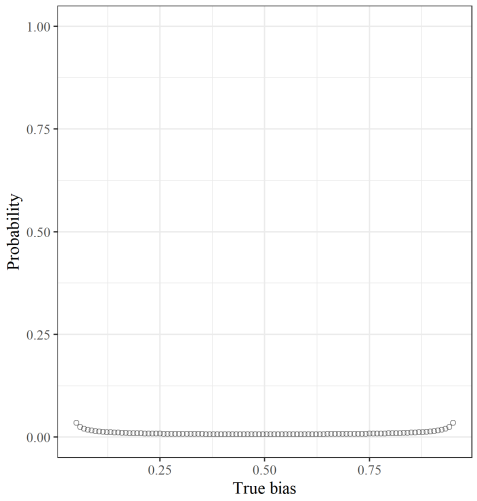

Figure 3: The actual probability of each value of the bias.

to the prior distribution is no guarantee of ecological optimality. Although this may seem an obvious point, its impact is often underestimated. Suppose the actual probability of each of the possible biases follows the distribution shown in Figure 3. As is clear from the figure, the probability distribution is very nearly flat, except that values of the bias that are on the extremes of the scale are somewhat more probable than values in the middle of the scale.[15] In other words, our flat prior distribution reflects the objective probability distribution very well, but it is slightly misspecified. Obviously, the objective distribution rather than our prior is what is going to determine which of the quadratic and Bayesian posteriors will be more ecologically rational. So which posterior distribution has the higher expected accuracy if we use the objective probability distribution to calculate the expectation? The answer is the quadratic posterior, which has an estimated expected inaccuracy of approximately 0.057 whereas the Bayesian posterior has an expected inaccuracy of approximately 0.06.

Obviously, we typically cannot know the objective probabilities of the hypotheses we are considering (assuming that the hypotheses have objective probabilities in the first place), but the point is this: unless we know that our prior probability function tracks the objective probability distribution extremely closely, we do not have good reason to think that strategies that maximize expected utility relative to our prior probability distribution are going to be successful in the actual world, since (as this example shows) even slight misspecification of the prior can render a strategy ecologically suboptimal.

[15] Mathematically, the distribution is proportional to $\frac{1}{\sqrt{(b(1-b))}}$, where b is the true bias of the coin.

The traditional Bayesian response to this issue is to perform a sensitivity analysis, where the dependence of the optimal decision on minor perturbations to the prior distribution is studied. However, an arguably equally rational response is to evaluate the various possible acts by a contextually appropriate and targeted decision criterion that we have reason to think will be more robust than expected utility maximization. For instance, a strong point that favors the Brier posterior over the Bayesian posterior in the example we have been considering is that it is more accurate for values of the bias close to 0 or 1. If we value accuracy on those regions of the probability scale, then we have a strong reason to favor the quadratic posterior, even though the Bayesian posterior has a slightly greater expected utility relative to our flat prior.

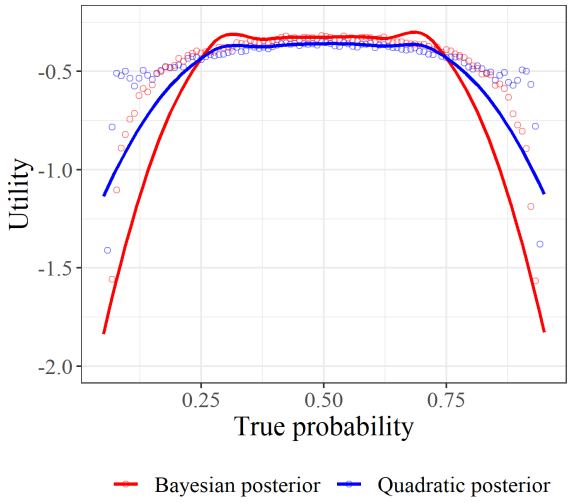

Figure 4: The utility of the Bayesian posterior predictive distribution and the quadratic posterior predictive distribution for each possible value of the parameter being estimated. The prior distribution is uniform and 20 data points generated from the true distribution are used for updating each distribution. Each data point is the result of averaging 1000 simulations.

But suppose we decide to go with expected utility maximization anyway. If we do, it is important that the utility function we use reflects our actual values, and the accuracy measure I have been using so far in this section arguably does not, because it implicitly treats an error on any part of the probability scale as equally severe, but we started the discussion of this example by assuming that errors on the extreme ends of the scale are worse. So let us consider an alternative utility function that better reflects our values. The following utility function serves as an example: $-|\mathrm{E}_{p_E}[H] - b|/(b(1-b))$ (where higher values are now better than lower ones). Figure 4 shows the utility of each posterior distribution at each possible value of the true bias, given that the posteriors have been updated on 20 data points.

It is clear from Figure 4 that the quadratic posterior has a much higher utility for extreme values of the bias, whereas the distributions have a similar utility in the middle of the range (because the utility function downplays mistakes in that range). We can also calculate an estimate of the *expected* utility of using either the Bayesian posterior or the quadratic posterior given that 20 data points are observed by calculating the average

utility of each posterior over all possible values of the bias in Figure 4. For the quadratic posterior, the resulting expected utility is roughly -0.5 and for the Bayesian posterior the expected utility is roughly -0.6, so the quadratic posterior has a higher expected utility than the Bayesian posterior.

This result seems to fly in the face of Greaves and Wallace's (2006) aforementioned result, but closer inspection reveals that there is no conflict . Greaves and Wallace's result establishes that the Bayesian posterior distribution maximizes expected utility given that the utility function is "proper", where a utility function is proper if and only if the expected utility of p relative to some distribution q, i.e., $\sum_i q(S_i) * U(S_i, p)$, is maximized by setting $q = p$. The utility function used in Figure 4 is not proper, and this is sufficient to explain why it is possible for the quadratic posterior to have a higher expected utility than the Bayesian posterior.

Propriety is a reasonable requirement to make of utility functions in certain contexts. For example, if the true hypothesis is one of the hypotheses under consideration and the incoming evidence is independent, then—as the amount of evidence increases—the average utility of the true hypothesis will be maximal only if the utility function is proper. Consequently, if we use our utility function to compare different hypotheses under consideration in the light of evidence, then—in the limit of increasing evidence—we will successfully identify the true hypothesis only if the utility function we use is proper. Hence, it makes sense to use a proper utility function as the basis of any updating rule, since only then will the updating rule be able to correctly identify the true hypothesis given increasing amounts of evidence. The quadratic posterior and the associated quadratic updating rule jointly satisfy this requirement since the Brier score is, indeed, a proper utility function.

But the fact that proper utility functions arguably are desirable for some purposes does not mean that improper utility functions are always inappropriate. An important distinction must be made between using a utility function for optimization purposes as opposed to using it purely for evaluative purposes. Because the utility function used in Figure 4 is improper, it is not a good idea to use it as the basis for an updating rule—that is, it is not a good idea to plug it in for the second term in 2.4 and solve the resulting optimization problem. The resulting updating rule is likely to perform quite poorly. But that does not mean that the utility function in Figure 4 is somehow illegitimate for assessing the utility of posterior distributions that have been generated via theoretically sound updating rules.

An analogy may be helpful. Suppose we are estimating the mean, μ of a normal distribution. There are strong statistical reasons for using the maximum likelihood estimate, which in this case will be the estimate that minimizes *squared* distance from data. However, given that we have an estimate in hand, the absolute distance is a perfectly sensible—and indeed more readily interpretable—measure of the accuracy of our estimate. In this case, the squared distance is the correct measure for optimization purposes, but the abso-

lute distance is a better measure for evaluation purposes. Similarly, even though the utility function used in Figure 4 arguably should not be used to select an estimate, it is a perfectly sensible utility function to use given that we have an estimate in hand and care more about being accurate for values of the bias close to 0 or 1. Indeed, if we insist on always using proper scoring rules to evaluate the epistemic utility of our probability functions, we will be severely limited in what values we are allowed to have.

In conclusion, even if we know that one of the hypotheses under consideration is true, Bayesian conditionalization may still not be the rationally optimal updating method from either an ecological or internal perspective. If the goal is to accurately estimate some probability, and it is important to have an accurate estimate given that the value of the probability is extreme, then there are good reasons for thinking that an updating method based on the Brier score rather than on the logarithmic score will do better. Greaves and Wallace's (2006) result implies that the Bayesian posterior is guaranteed to have a higher expected utility if utility is measured with a utility function that is proper, but we have seen reasons for thinking that this result is less impressive than it initially seems: first, the argument hinges on measuring utility with a proper utility function, which arguably does not adequately reflect the actual utilities in the example; second, if maximizing ecological rationality is a priority, then the expected utility calculated relative to an agent's subjective probability distribution is of limited value, unless the agent's distribution is known to be close to the actual probability distribution (assuming such a distribution exists).

The argument in this section is admittedly weaker than the argument in the preceding section, because whereas the stacking posterior implements a well-known state-of-the art prediction averaging method that has proven its worth in many real-world applications, the "quadratic posterior" that I suggest in this section has to my knowledge never been suggested in the published literature. It therefore remains to be seen whether the merits of quadratic updating extend beyond the very simple kind of example we have considered in this section.

5 When is Bayesian conditionalization guaranteed to be ecologically optimal?

The examples in the preceding sections show that there are theoretically well-motivated updating rules that are arguably more rational—both ecologically and from the agent's own point of view—than Bayesian conditionalization, at least under certain conditions. In fact, we can use the examples from those sections to demarcate more precisely when we might expect Bayesian conditionalization to be *optimal*. In particular, if any of the following conditions are violated, then Bayesian conditionalization is not guaranteed to be

ecologically optimal:

1. The hypotheses under consideration form a partition.

2. The agent's joint distribution over the hypotheses and evidence accurately reflects the objective probability distribution, if one exists.

3. The agent's goal is to maximize expected utility with respect to a proper utility function.

To see why satisfying each of (1)-(3) is necessary for Bayesian conditionalization to be guaranteed to be ecologically optimal, note that the example in Section 3 violates (1) and (2). Similarly, the example in Section 4 shows that Bayesian conditionalization is not necessarily ecologically optimal if either (2) or (3) is violated. Therefore, (1)-(3) are necessary conditions for conditionalization to be guaranteed to be ecologically rational.

Is satisfying (1)-(3) *sufficient* to guarantee that Bayesian conditionalization will be ecologically optimal? There is reason to think that the answer is yes. First, if (1) and (3) are satisfied, then the arguments against using expected utility maximization that were presented in sections 3 and 4 do not go through. Furthermore, on the condition that expected utility maximization is performed with a proper utility function, we know from Greaves and Wallace (2006) that conditionalization will be an optimal act.

Hence, there are good reasons for thinking that (1)-(3) are jointly necessary and sufficient for Bayesian conditionalization to be guaranteed to be ecologically optimal. Of course, one may question the importance of such a finding, since a strategy need not be ecologically optimal in order to be ecologically rational for an agent—if the strategy effectively achieves the agent's goals, then it is arguably rational even if it fails to be optimal. My guess is that Bayesian conditionalizations and approximations thereof may indeed often be adequate even in cases where they fail to be optimal. For example, it's clear from Figure 2 that the difference between Bayesian conditionalization and quadratic updating is probably not going to be significant in many contexts.

However, I think the findings in this section should still be of interest, because it is common for Bayesian epistemologists to think that Bayesianism is not just one option among many, but that it instead is the uniquely rational option regardless of agents' values or epistemic situation. The findings in this section push heavily against such a conception.

6 Concluding remarks

The paper has argued that Bayesian conditionalization fails to be ecologically optimal and internally rational in scenarios that are arguably quite common in practice. Along

the way, we have seen reasons to be skeptical of the utility of common decision making rules, especially expected utility maximization. We have also seen that the conditions under which Bayesian conditionalization is guaranteed to be optimal are quite narrow. The standard picture, according to which Bayesian conditionalization is the uniquely rational updating rule, is false—even for ideal reasoners who face no resource constraints.

Readers may worry that the alternative picture that we end up with is hopelessly messy and context-sensitive, where nothing of any generality can be said about how rational agents should respond to evidence (cf. Carr (2021)). I think there is good reason to think that this will not be the case. The minimum divergence metaframework gives us a principled way of evaluating the cogency of probabilistic belief updating rules in specific contexts. This metaframework may in turn be regarded as a specific implementation of an "optimization" approach to epistemology, which holds that epistemic strategies should be evaluated by seeing whether they optimize some given epistemic target (cf. Schurz (2021), who argues for an "optimization" framework for epistemic justification and discusses several applications). There is therefore reason to think there is plenty of middle ground between an "ideal" monolithic conception of rationality and a radically context-sensitive one.

References

Alchourrón, C., Gärdenfors, P., and Makinson, D. (1985). On the Logic of Theory Change: Partial Meet Contraction and Revision Functions. *J. Symb. Log.*, 50:510–530.

Bacchus, F., Kyburg, H., and Thalos, M. (1990). Against Conditionalization. *Synthese*, 85:475–506.

Berger, J. O., Bernardo, J. M., and Sun, D. (2009). The Formal Definition of Reference Priors. *The Annals of Statistics*, 37(2):905–938.

Bernardo, J. M. (1979). Reference Posterior Distributions for Bayesian Inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):113–147.

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

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

Bona, G. D. and Staffel, J. (2017). Graded Incoherence for Accuracy-Firsters. *Philosophy of Science*, 84(2):189–213.

Breiman, L. (1996). Stacked regressions. *Machine Learning*, 24(1):49–64.

Carr, J. R. (2021). Why Ideal Epistemology? *Mind*.

De Bona, G. and Staffel, J. (2018). Why be (approximately) coherent? *Analysis*, 78(3):405–415.

Diaconis, P. and Zabell, S. L. (1982). Updating Subjective Probability. *Journal of the American Statistical Association*, 77(380):822–830.

Douven, I. (2013). Inference to the Best Explanation, Dutch Books, and Inaccuracy Minimisation. *The Philosophical Quarterly*, 63(252):428–444.

Douven, I. (2016). Explanation, Updating, and Accuracy. *Journal of Cognitive Psychology*, 28(8):1004–1012.

Douven, I. (2020). The ecological rationality of explanatory reasoning. *Stud Hist Philos Sci*, 79:1–14.

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(2):535–570.

Easwaran, K. (2021). *An Opinionated Introduction to the Philosophical Foundations of Bayesianism*.

Eva, B. and Hartmann, S. (2018). Bayesian Argumentation and the Value of Logical Validity. *Psychological Review*, 125(5):806–821.

Forster, M. R. and Sober, E. (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., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press, Boca Raton, FL, third edition.

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

Gigerenzer, G. and Todd, P. M. (2012). *Ecological Rationality: The Normative Study of Heuristics*. Oxford University Press.

Greaves, H. (2013). Epistemic Decision Theory. *Mind*, 122(488):915–952.

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

Grünwald, P. and van Ommen, T. (2017). Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. *Bayesian Analysis*, 12(4):1069–1103.

Holmes, C. C. and Walker, S. G. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104(2):497–503.

Kullback, S. and Leibler, R. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

Marsh, B., Todd, P. M., and Gigerenzer, G. (2004). Cognitive Heuristics: Reasoning the Fast and Frugal Way. In *The nature of reasoning.*, pages 273–287. Cambridge University Press, New York, NY, US.

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press, Boca Raton, FL.

Pettigrew, R. (2021). On the pragmatic and epistemic virtues of inference to the best explanation. *Synthese*.

R Core Team (2020). R: A Language and Environment for Statistical Computing.

Salmon, W. C. (1990). The Appraisal of Theories: Kuhn Meets Bayes. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1990:325–332.

Schurz, G. (2021). Optimality justifications and the optimality principle: New tools for foundation-theoretic epistemology. *Nous*, n/a(n/a).

Selten, R. (1998). Axiomatic Characterization of the Quadratic Scoring Rule. *Experimental Economics*, 1:43–62.

Shimony, A. (1970). Scientific inference. In Colodny, R., editor, *The Nature and Function of Scientific Theories*, page 4. University of Pittsburgh Press.

Sober, E. (2008). *Evidence and Evolution: The Logic Behind the Science*. Cambridge University Press.

Spohn, W. (2012). *The Laws of Belief: Ranking Theory and Its Philosophical Applications*. Oxford University Press, Oxford.

Sprenger, J. (2019). Conditional Degree of Belief and Bayesian Inference. *Philosophy of Science*, 87(2):319–335.

Stanford, K. P. (2006). *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives*. Oxford University Press, New York, NY.

Sterkenburg, T. F. and Grünwald, P. D. (2021). The no-free-lunch theorems of supervised learning. *Synthese*.

Todd, P. and Brighton, H. (2015). Building the Theory of Ecological Rationality. *Minds and Machines*, 26:9–30.

Tversky, A. and Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124.

van Fraassen, B. C. (1981). A Problem for Relative Information Minimizers in Probability Kinematics. *British Journal for the Philosophy of Science*, 32(4):375–379.

Vassend, O. B. (2018). Goals and the Informativeness of Prior Probabilities. *Erkenntnis*, 83(4):647–670.

Vassend, O. B. (2019). Justifying the Norms of Inductive Inference. *British Journal for the Philosophy of Science*.

Vranas, P. B. M. (2000). Gigerenzer's normative critique of Kahneman and Tversky. *Cognition*, 76(3):179–193.

Williams, P. M. (1980). Bayesian Conditionalisation and the Principle of Minimum Information. *The British Journal for the Philosophy of Science*, 31(2):131–144.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2):241–259.

Wolpert, D. H. (1996). The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation*, 8(7):1341–1390.

Yao, Y., Aki, V., Daniel, S., and Andrew, G. (2018). Using Stacking to Average Bayesian Predictive Distributions (with Discussion). *Bayesian Analysis*, 13(3):917–1007.

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

Zynda, L. (1996). Coherence as an ideal of rationality. *Synthese*, 109(2):175–216.