EUCLIDEAN PROPERTIES OF BAYESIAN UPDATING

KYLE P. CHAUVIN

December 12, 2021

Abstract. This paper introduces a novel framework for analyzing Bayesian updating and non-Bayesian heuristics. A *learning rule* consists of an arbitrary set of *belief* states and a set of transition functions, called *arguments*, from the belief set to itself. Bayesian learning rules, with beliefs in the form of probability distributions over a state and arguments in the form of Bayes' rule, are a special case. The paper's first main result is an axiomatic characterization of *Virtual Bayesian* learning rules, which can be turned into a Bayesian via a relabeling of the set of beliefs. There are three substantive axioms – that arguments are injective functions, that arguments commute, and that repeated application of an argument never produces cycles of beliefs – as well as three regularity assumptions. The axioms both identify the algebraic properties common to all Bayesians and distinguish which among familiar updating heuristics are Virtual Bayesians. The second main result establishes that any Virtual Bayesian learning rule can be embedded into Euclidean space – and therefore equipped with geometric notions of magnitude, direction, etc. – by defining the 'agreement' between pairs of arguments in a suitably additive manner. Applying such an embedding, an argument's direction corresponds to the limit of the support of posterior beliefs under repeated application of the argument, and its magnitude is the extent to which a single application pushes prior beliefs towards that limit. The paper discusses how the framework of learning rules could be applied to additional contexts, including the elicitation of beliefs from laboratory subjects.

JEL Classification: D83, C11, C44

Keywords: Non-Bayesian Updating, Automata, Group Theory, Information Measurement

NYU Shanghai Economics; kyle.chauvin@nyu.edu. My thanks to Roland Bénabou, Emiliano Catonini, Ben Golub, Faruk Gul, Wolfgang Pesendorfer, Matthew Rabin, Tomasz Strzalecki, Michael Thaler, Nikhil Vellodi, Leeat Yariv, and Huseyin Yildirim for helpful comments.

1. Introduction

Departures from Bayesian updating take various forms. An agent may learn according to an incorrectly specified but otherwise standard Bayesian model,[1] or she may employ a non-Bayesian heuristic characterized by an intuitive functional form (e.g. DeGroot learning) or by axiomatic properties.[2] In this paper, I analyze Bayesian updating as well as non-Bayesian heuristics with the following framework patterned after concepts in group theory and automata theory. Given an arbitrary set of beliefs $\mathcal{B}$, an *argument* is any function $a : \mathcal{B} \longrightarrow \mathcal{B}$, representing how the agent's beliefs should update as a result of receiving the argument. Arguments compose with each other to form new arguments; for example, the composition $a_1 \circ a_2$ is $b \longmapsto a_1(b) \longmapsto a_2(a_1(b_1))$. A *learning rule* is a pair $(\mathcal{A}, \mathcal{B})$, where $\mathcal{A}$ is a set of arguments over $\mathcal{B}$ which is closed under composition.

This paper is structured around two central questions. The first asks which algebraic properties of arguments and beliefs characterize Bayesian learning rules. Motivated by the fact that the answer involves interpreting Bayesian arguments as vectors which 'push' beliefs in different directions, the second question asks which other learning rules admit geometric representations. The applicability of the answers to both questions is illustrated in the context of belief elicitation in the lab. Finally, I discuss how additional applications can leverage the framework of learning rules to complement the traditional probabilistic perspective to belief updating.

The paper's first central question is, formally: under what conditions can a learning rule be transformed into a Bayesian by relabeling its belief set? When this is possible, the learning rule is termed *Virtual Bayesian.* For example, consider how someone might update about the risk of her house flooding in the next five years. Initially, she expresses her belief in the form of the verbal statement, 'the risk is small.' Then it rains heavily some month, and she thinks, 'there's some risk.' After half a year of minimal rain she again believes, 'the risk is small,' but in the following month a downpour nearly floods her house, after which she maintains, 'the risk is real.' The woman's verbal statements are non-probabilistic beliefs, and her updating from one belief to another is not governed by Bayes' rule. Nonetheless, if we relabel her beliefs as follows:

[1] See, for instance, Rabin and Schrag (1999); Eyster and Rabin (2010); Esponda and Pouzo (2016).

[2] See Epstein (2006); Lehrer and Teper (2016); Cripps (2019) and others discussed at the end of this section.

$$\left\{\begin{array}{rcl}\text{`the risk is small'} & \Longrightarrow & \text{`the house floods with probability 1\%'}\\ \text{`there's some risk'} & \Longrightarrow & \text{`the house floods with probability 3\%'}\\ \text{`the risk is real'} & \Longrightarrow & \text{`the house floods with probability 8\%'}\end{array}\right.$$

and specify the following conditional probabilities of monthly rain level given flood propensity:

	P conditional on:	
	flooding	no flooding
rains heavily	12.67%	4.14%
minimal rain	78.72%	94.86%
almost floods	8.61%	1.0%

then the woman resembles a Bayesian: each transition from one belief to another follows the functional form of Bayes' rule using the above conditional probabilities. Hence, she is a Virtual Bayesian.

Now consider a man who updates about flood risk in a different manner. When it rains in a given month, he assesses the risk of future flooding as 'high.' When it does not rain, he assesses the risk as 'low.' In contrast to the woman, there is no way to relabel his beliefs so he becomes Bayesian. If there were, then – under the relabeling – he would learn by adjusting weight on some set of underlying states via Bayes' rule. The information conveyed to him by a rainy month would cause him to place greater weight on *at least one* of his underlying states, and therefore two months of rain would lead to even greater weight on that state. However, after two rainy months the man instead retains exactly the same belief he had after one rainy month. Hence, he is not a Virtual Bayesian.

Theorem 1 axiomatically characterizes when a learning rule $(\mathcal{A}, \mathcal{B})$ is a Virtual Bayesian. There are three substantive axioms: that arguments commute, that arguments are injective, and that arguments do not produce cycles of beliefs. Commutativity means that an agent is insensitive to the order in which arguments arrive. Injectivity means different priors lead to different beliefs or, equivalently, that after any sequence of arguments and given an agent's terminal belief, there is no ambiguity as to where she started. Acyclicality ensures that repeatedly applying any argument leads the agent to new posterior beliefs. There are also three regularity axioms: that there are only countably infinite arguments and beliefs, that there exists an 'original' prior belief from which all beliefs can be reached by way of a unique argument, and that there are at least two 'directionally distinct' arguments for which no positive multiple of the first coincides with a positive multiple of the second.

It is straightforward to verify that Bayesian learning rules satisfy the above properties; the challenge is to start with an arbitrary learning rule $(\mathcal{A}, \mathcal{B})$ that satisfies them and to produce a matching Bayesian. The proof of Theorem 1 proceeds by augmenting the set of arguments to include inverses and fractional elements, forming a vector space over the rational numbers. This enriched copy of $\mathcal{A}$ is then embedded inside the real line, associating each original argument with a value that is re-interpreted as a log likelihood-ratio in a Bayesian learning rule.

Virtual Bayesians take various forms. Distorted copies of Bayes' rule, such as probability weighting, are familiar examples. Other families of updating on probabilities, for example power-law updating of the form $p \longmapsto p^x$, are also Virtual Bayesians. Some heuristics studied in the learning-in-games literature, such as reinforcement learning and fictitious play, encode Virtual Bayesian learning rules despite their patently non-Bayesian configurations. Further afield, there are cases of beliefs in the form of verbal statements, or forms of heuristic thinking, such as pro-con lists and multi-cell rubrics, which satisfy all six axioms of Theorem 1. There are also seemingly equally plausible heuristics which fail one or more of the axioms. For instance, any finite learning rule cannot satisfy acyclicality; learning rules that combine two arguments by preserving the most compelling one and discarding the other are necessarily not injective; combining arguments by averaging with fixed weights, ala DeGroot learning, necessarily violates commutativity. 'Deductive' learning rules, which transition among beliefs by ruling out potential values of an underlying state, are commutative but neither injective nor acyclic. These examples illustrate how Theorem 1's axioms can disentangle Virtual Bayesians from non-Bayesians.

The paper's second central question asks: when and how can the arguments of a learning rule be be endowed with geometric notions of direction and magnitude? The answer starts by considering how to define an 'agreement' function $\mathcal{A} \times \mathcal{A} \longrightarrow \mathbf{R}$. Such a function is *additive* if the agreement between any two arguments a_1 and a_2 is unchanged when either or both of them is decomposed into component parts. Functions which satisfy this property and two regularity conditions are the foundation for embedding the set of arguments into $\mathbf{R}^n$. As Theorem 2 establishes, it is possible to define an additive agreement function on any Virtual Bayesian, and any learning rule which admits such a function and satisfies the regularity axioms of Theorem 1 is necessarily a Virtual

Bayesian. Theorem 3 makes explicit the link between additive agreement functions and embeddings in Euclidean space: for every additive agreement function on a Virtual Bayesian, there is a unique embedding into $\mathbf{R}^n$ that extends the additive agreement function to the standard inner product.

Viewed as subspaces of $\mathbf{R}^n$, Virtual Bayesian learning rules are endowed with an *agreement geometry*. Notions of norm, angle, and projection all follow from the associated additive agreement function. As illustrated through a series of examples, an agreement geometry describes the effect that an argument has on an agent's prior belief in a way that complements more familiar descriptions. Every argument is characterized by its magnitude and direction. For a Bayesian learning rule, each direction (outside of a measure-zero set) corresponds to a unique value of the underlying latent state variable. Positively extending an argument while holding its direction fixed produces an exaggerated version which, in the limit, sends all prior beliefs arbitrarily close to the direction's associated state value. Moreover, the norm defined on the set of arguments constitutes a *prior-free* measure of the strength of information. In contrast, quantifying information via entropy reduction or by value gained in a decision problem (as characterized in Frankel and Kamenica (2019)) necessarily depends on one's prior belief. As an example of the wedge between the two approaches, if one's prior is the uniform distribution, then for any fixed direction, an argument of higher magnitude yields a posterior with more entropy reduced; however, if one's prior places almost all weight on a single state, higher magnitude arguments pushing towards alternative states induce *more* entropy.

To illustrate the tools developed in this paper, I consider how to use the characterization of Virtual Bayesians to elicit laboratory subjects' beliefs. Standard methods of elicitation (see Schotter and Trevino (2014) and Schlag, Tremewan and van der Weele (2015) for literature reviews) typically make assumptions about subjects' preferences and then leverage subjects' decisions over lotteries to deduce their beliefs.[3] By contrast, I propose a class of methods that link the context of the subject's belief to a data-generating procedure in the lab and then present the subject with hypothetical additional information. For a subject known to be a Virtual Bayesian who can

[3] For example, quadratic and other proper scoring rules make truthful reporting of one's beliefs incentive compatible under the assumption of risk neutrality. Various papers have developed methods for de-biasing beliefs skewed by probability weighting or other biases. However, these methods are all predicated on subjects possessing *some* subjective probability distribution over outcomes.

describe his beliefs in rudimentary geometric terms (e.g. distinguishing relative angle or relative magnitudes), his reaction to additional information is sufficient to identify his prior belief.

This paper fits into an emerging literature that axiomatically characterizes non-Bayesian updating. Most closely related is Cripps (2019), which also characterizes an agent who becomes a Bayesian updater when his beliefs are properly translated. However, the agent in Cripps' setup maintains beliefs in the form of probability distributions over an explicit state, and he receives information through statistical (Blackwell) experiments. Thus while Cripps' agent can be translated into a Bayesian with fewer axioms,[4] the agent starts out already closer to the Bayesian framework than an abstract learning rule. In a related vein, Shmaya and Yariv (2016) characterize when *trees* of probabilistic beliefs can be rationalized as conditional expectations processes. As they allow for history-dependent correlation structures, Shmaya and Yariv obtain a more permissive condition for near-Bayesianness: that a belief at any one node must lie in the convex hull of beliefs at successor nodes. Zhao (2016) provides a way to *augment* Bayesian updating procedures so that an agent can process information of the form 'event A is more likely than event B.'

In the choice theoretic paradigm, Epstein (2006) presents a generalized version of the Anscome-Aumann theorem to allow for dynamically inconsistent updating, and Epstein, Noor and Sandroni (2008) extends this to a repeated context. Epstein and Seo (2010) presents a generalization of the de Finetti theorem in which agents are not certain that successive signals are conditionally independent. More recently, Hanany and Klibanoff (2014) characterizes dynamic consistency in the context of ambiguity aversion, and Lehrer and Teper (2016) studies agents who satisfy a weaker version of Bayesianism called 'local consistency.' Although I am unaware of an equivalent definition of learning rules in the economics literature, automata have been used as a model of limited cognition. For example (Abreu and Rubinstein, 1988) study repeated games with strategies implemented via finite automata, and Wilson (2014) characterizes how finite state machines can optimally approximate Bayesian learning in dynamic environments. Similarly, although the concept of agreement geometry is novel, conceiving of and manipulating information as log-likelihood-ratio vectors finds precedence in a variety of papers, e.g. Molavi, Tahbaz-Salehi and Jadbabaie (2018).

[4] The key characterizing axiom in Cripps (2019) is *divisibility*: updating from two pieces of information produces the same posterior as sequentially updating from each of them one at a time. This corresponds to commutativity (A.5) in the present paper. Additionally, Cripps' non-dogmatic axiom contains a requirement of injectivity. His other two axioms are more closely tied to the setup of his model and do not have close counterparts in the present paper.

Organization. Section 2 introduces the machinery of learning rules, defines Bayesianism in this context, and lays the foundation for the main characterization theorem presented in Section 3. Following several examples of Virtual and non-Bayesians, Section 4 analyzes the geometric structures of Virtual Bayesians. Section 5 uses the paper's results to construct several belief elicitation methods. The discussion in Section 6 considers a series of open questions to guide future applications of the Virtual Bayesian framework. All proofs omitted from the main text are found in Appendix A.

2. Model: Learning Rules

Let $\mathcal{B}$ denote an arbitrary set. Elements $b \in \mathcal{B}$ are called belief states, and $\mathcal{B}$ is called a belief set. An ***argument*** over $\mathcal{B}$ is a function $a : \mathcal{B} \longrightarrow \mathcal{B}$. An argument describes how a learning agent's beliefs should change as a result of receiving the argument. The composition of two arguments a_1 and a_2 is defined by $a_1 \circ a_2 : b \longmapsto a_2(a_1(b))$. The identity argument $b \longmapsto b$, which communicates no information, is denoted a_{id}. By pairing a set of arguments with its corresponding belief set we obtain a simple model of updating.[5]

Definition. *A **learning rule** $(\mathcal{A}, \mathcal{B})$ is set of arguments $\mathcal{A}$ over belief set $\mathcal{B}$ such that $\mathcal{A}$ is closed under composition: $a_1 \circ a_2 \in \mathcal{A}$ for all $a_1, a_2 \in \mathcal{A}$.*

To illustrate this concept, first consider several examples of updating as described in the language of learning rules. Readers eager for the formal definition of Bayesian vs non-Bayesian learning rules can skip to Section 2.1.

Agent 1. *The Bernoulli Bayesian.* A Bayesian agent has uncertainty about the bias of a coin. She believes there are two equally likely states, a heads-biased state in which the probability of flipping heads is $q \in (1/2, 1)$ and a tails-biased state in which heads realizes with probability $1 - q$. She learns about the coin by observing realizations of i.i.d. flips. This agent's set of beliefs is made up of the conditional probability assessments she could have about the coin's bias after a finite number of observations: $\mathcal{B} = \{\mathbf{P}[\textsf{heads-biased} \mid y_0, \ldots, y_k]\} \subset (0, 1)$, where y_i denotes the realization of flip

[5] In algebraic terminology, argument set $\mathcal{A}$ constitutes a *semigroup*, a set with a composition operation which is closed under composition and satisfies the associative law $a_1 \circ (a_2 \circ a_3) = (a_1 \circ a_2) \circ a_3$. (Function composition always satisfies the associative law.) Note that $\mathcal{A}$ need not include the identity argument a_{id}, although many examples studied in the paper, including all Bayesians, do include a_{id}. Moreover, the pair $(\mathcal{A}, \mathcal{B})$, in which elements of $\mathcal{A}$ are 'acting' on $\mathcal{B}$, constitutes a *semigroup action.* Equivalently, $(\mathcal{A}, \mathcal{B})$ can be viewed as a (potentially non-finite) automaton in which elements of $\mathcal{B}$ describe the current state of abstract machine and arguments in $\mathcal{A}$ correspond to inputs to the machine which induce deterministic transitions among the states. For reference, see Reddy (2014) among others.

$i = 0, \ldots, k$. Arguments correspond to sequences of additional flips of the coin: given sequence $(\hat{y}_0, \ldots, \hat{y}_{\hat{k}})$, the associated argument is

$$a_{(\hat{y}_0,\ldots,\hat{y}_{\hat{k}})} : \mathbf{P}[\textsf{heads-biased} \mid y_0, \ldots, y_k] \longmapsto \mathbf{P}[\textsf{heads-biased} \mid y_0, \ldots, y_k, \hat{y}_0, \ldots, \hat{y}_{\hat{k}}].$$

Each argument is in the form of Bayes' rule. For example, a single additional heads flip $\hat{y} = \textsf{heads}$ corresponds to

$$a_{\textsf{heads}} : \pi \longmapsto \frac{q\pi}{q\pi + (1-q)(1-\pi)}$$

for all $\pi \in \mathcal{B}$. Any flip sequence with h net number of heads minus tails is associated with the argument

$$a_h : \pi \longmapsto \frac{q^h \pi}{q^h \pi + (1-q)^h (1-\pi)}.$$

This illustrates how all arguments correspond to a certain number of net heads flips. Moreover, the composition of an h_1 net-heads argument with an h_2 net-heads argument is

$$a_{h_1} \circ a_{h_2} = a_{h_1+h_2} : \pi \longmapsto \frac{q^{h_1+h_2}\pi}{q^{h_1+h_2}\pi + (1-q)^{h_1+h_2}(1-\pi)}.$$

In this way, the agent combines two pieces of information by adding together their net-heads values.

Agent 2. *The Beta Bayesian.* Another Bayesian learns about the bias of a coin by observing realizations of i.i.d. flips. However, he believes the probability of a heads flip is uniformly distributed on $(0, 1)$. After observing $y_0, \ldots, y_k$, with $\mathsf{h}(y_0, \ldots, y_k)$ total heads and $\mathsf{t}(y_0, \ldots, y_k)$ total tails, the agent believes the coin's bias is distributed $\textsf{Beta}(1 + \mathsf{h}(y_0, \ldots, y_k), 1 + \mathsf{t}(y_0, \ldots, y_k))$. This agent's belief set is therefore $\mathcal{B} = \{\textsf{Beta}(h, t) \mid h, t \geq 0\}$, and his set of arguments is

$$\mathcal{A} = \left\{a : \textsf{Beta}(h, t) \longmapsto \textsf{Beta}(h + h', t + t') \mid h', t' \geq 0\right\}.$$

He combines arguments a_{h_1,t_1} and a_{h_2,t_2} into $a_{h_1+h_2,t_1+t_2}$, that is by adding together the individual dimensions separately.

Agent 3. *The Flip-Flopper.* A non-Bayesian learns about a coin in the following manner. She thinks the coin always flips heads or always flips tails. After observing a heads flip, she concludes it always flips heads; after a tails flip, she thinks the coin always flips tails. Her belief set is $\mathcal{B} = \{\textsf{always heads}, \textsf{always tails}\}$, and her argument set comprises the two maps $a_h : b \longmapsto \textsf{always heads}$ and $a_t : b \longmapsto \textsf{always tails}$. For this agent, $a_t \circ a_h = a_h$ and $a_h \circ a_t = a_t$.

Agent 4. *The Verbal Reasoner.* Another non-Bayesian describes his uncertainty about a coin with a set of verbal statements which can be ordered in terms of their confidence in predicting heads flips. One statement is *the coin is neutral.* The next most heads-friendly statement is *maybe heads*, then *probably heads*, then *likely heads*, then *very likely heads*, *very very likely heads*, *very very very likely heads*, and so on. He likewise entertains the possibility of *maybe tails*, *probably tails*, and so forth. The agent's belief set $\mathcal{B}$ is the union of all possible statements he could make about the coin. When the agent observes a heads flip, his belief progresses one statement in the heads-friendly direction; a tails flip sends him one statement backwards. The agent's argument set $\mathcal{A}$ is the union of transitions $k \geq 0$ statements forwards and $j \leq 0$ statements backwards.

Commentary on the Examples. These four examples collectively illustrate several features of learning rules. First, learning rules can describe any case of Bayesian updating which is sufficiently stationary, that is in which the information content of a signal realization remains consistent across signals. Second, they can also describe a wide range of non-Bayesian updating examples. Most importantly, by framing updating in terms of belief-transition maps who compose with each other, learning rules provide a method for analyzing the compositional structure of learning models as distinct from particular functional forms. Defined rigorously in the following subsection, two learning rules are said to be *isomorphic* if the set of arguments in the first learning rule is equivalent to that in the second learning rule under a one-to-one relabeling of the belief states. Thus, some non-Bayesian updating rules, despite not encoding beliefs as probabilities or not transitioning among them via Bayes' rule, nonetheless possess a Bayesian compositional structure.

For example, Agent 4 is clearly not Bayesian. However, it is possible to relabel his set of beliefs so that he matches the Bayesian updating of Agent 1. Agent 4's statement *the coin is neutral* becomes $\pi = 1/2$. His statement *maybe heads* is mapped to $\pi = q$, *probably heads* to $\pi = q^2/(q^2 + (1-q)^2)$, and so forth. This mapping associates the verbal statement Agent 4 would adopt after every possible sequence of flip realizations with the corresponding belief that Agent 1 would adopt. Under this relabeling of belief states, Agents 1 and 4 have the same set of arguments; each argument in each learning rule simply pushes the corresponding agent's beliefs forwards or backwards along a single dimension. Agent 4 is accordingly termed a *Virtual Bayesian.*

In constrast, Agent 3 cannot be paired with any Bayesian, as her learning rule contains several features which no Bayesian does. First, the way she composes arguments displays order dependence:

$a_t \circ a_h \neq a_h \circ a_t$. Next, her arguments not injective, e.g. $a_t(\textsf{always heads}) = a_t(\textsf{always tails})$. Finally, her learning rule contains *cyclic* arguments for whom multiple application is equivalent to a single application: $a_t \circ a_t = a_t$ and $a_h \circ a_h = a_h$. As is established formally in the following section, no Bayesian learning rule has these properties, and therefore there is no Bayesian with which Agent 3 is isomorphic.

Finally, it is worth comparing Agents 1 and 2. Both of them have Bayesian learning rules, but they are not isomorphic. Agent 1 is concerned with the net number of heads flips she has observed. She considers two arguments with x_1 and x_2 net heads flips as equivalent to a single argument that presents $x_1 + x_2$ net heads flips. Her set of arguments is thus isomorphic to the additive group of integers. On the other hand, Agent 2 keeps track of *both* the total number of heads observed and the total number of tails observed. For example, he distinguishes between the sequences (H, H, T) and (H, H, H, T, T), whereas Agent 1 would consider those to be identical arguments. His argument set is isomorphic to the (algebraically distinct) additive semigroup $\mathbf{Z}_{>0} \times \mathbf{Z}_{>0}$.

2.1. Bayesian and Virtual Bayesian Learning Rules

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $\{X, (Y_i)_{i=1}^{\infty}\}$ be a set of random variables. The 'state' X has support $\mathcal{X}$, and the 'signals' Y_i share a common support $\mathcal{Y}$. After observing the sequence $y_1, \ldots, y_n$, a Bayesian learner's belief is the conditional distribution $\mathbf{P}_X[\cdot|y_1, \ldots, y_n]$.

Several regularity assumptions are placed on this environment.[6] First, assume that the Y_i's are *i.i.d.* conditional on X, which ensures that the learning environment is stationary. Next, the following conditions ensure that Bayes' rule is always applicable and unambiguously defined after any sequence of $\mathcal{Y}_i$ realizations: (1) $\mathcal{Y}$ is discrete, (2) $\mathcal{X}$ is either discrete or continuous (for notational purposes I use the former), and (3) $\mathbf{P}[Y_i = y\,|x] > 0$ for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$. Finally, assume that the learning environment is non-trivial: there exists some $y \in \mathcal{Y}$ such that $\mathbf{P}_X[\cdot|y] \neq \mathbf{P}_X[\cdot]$.

Definition. *The* ***Bayesian*** *learning rule corresponding to* $\{X, (Y_i)_{i=1}^{\infty}\}$ *combines the belief set*

$$\mathcal{B} = \{\mathbf{P}_X[\cdot|\ y_1, \ldots, y_n] \mid n \geq 0\}$$

[6] Section 6 includes further discussion about the significance of these conditions.

with the argument set

$$\mathcal{A} = \{\mathbf{P}_X[\cdot|\ y_1, \ldots, y_n] \longmapsto \mathbf{P}_X[\cdot|\ y_1, \ldots, y_n, \hat{y}_1, \ldots, \hat{y}_m] \mid n, m \geq 0\}\,.$$

Per Bayes' rule, each argument $a = a_{(\hat{y}_1,\ldots,\hat{y}_m)}$ has the functional form

$$a_{(\hat{y}_1,\ldots,\hat{y}_m)}(\pi)(x) = \frac{\pi(x) \cdot \mathbf{P}[\hat{y}_1|x] \cdot \cdots \cdot \mathbf{P}[\hat{y}_m|x]}{\sum_{x' \in \mathcal{X}} \pi(x') \cdot \mathbf{P}[\hat{y}_1|x'] \cdot \cdots \cdot \mathbf{P}[\hat{y}_m|x']}$$

for $\pi \in \mathcal{B}$ and $x \in \mathcal{X}$.

Morphisms. To formally define the concept of Virtual Bayesians, whose compositional structures match those of Bayesians, I first clarify the definition of homomorphisms (structure-preserving maps) and isomorphisms (bijective homomorphisms) between learning rules. Note that these are entirely standard definitions; they coincide with semigroup action homomorphisms and isomorphisms.

Definition. *A* ***homomorphism*** *from $(\mathcal{A}, \mathcal{B})$ to $(\mathcal{A}', \mathcal{B}')$ is a pair of maps $g : \mathcal{A} \longrightarrow \mathcal{A}'$ and $h : \mathcal{B} \longrightarrow \mathcal{B}'$ such that $g(a)(h(b)) = h(a(b))$ for all $a \in \mathcal{A}$ and $b \in \mathcal{B}$. An isomorphism is a homomorphism whose maps g and h are both invertible.*

Definition. *Learning rule $(\mathcal{A}, \mathcal{B})$ is a* ***Virtual Bayesian*** *if it is isomorphic to some Bayesian learning rule $(\mathcal{A}', \mathcal{B}')$.*

There is also an equivalent and perhaps more intuitive condition for $(\mathcal{A}, \mathcal{B})$ and $(\mathcal{A}', \mathcal{B}')$ to be isomorphic: for some invertible map $h : \mathcal{B} \longrightarrow \mathcal{B}'$,

$$\mathcal{A} = \{b \longmapsto h^{-1}(a'(h(b))) \mid a' \in \mathcal{A}'\}.$$

That is, each argument $a \in \mathcal{A}$ is equivalent to a unique counterpart $a' \in \mathcal{A}'$ when the elements of $\mathcal{B}$ are 'relabeled' through h into elements of $\mathcal{B}'$.

3. Algebraic Properties of Virtual Bayesians

This section establishes the paper's main result, an axiomatic characterization in algebraic terms of Virtual Bayesian learning rules. Following the proof of Theorem 1, I discuss a series of examples of Virtual Bayesians and non-Bayesians.

Theorem 1. *Learning rule $(\mathcal{A}, \mathcal{B})$ is a Virtual Bayesian if and only if it is*

(A.1) ***countable:*** *$\mathcal{A}$ and $\mathcal{B}$ are countably infinite,*

(A.2) ***self-recording:*** *there exists $b_0 \in \mathcal{B}$ such that, for all $b \in \mathcal{B}$, there exists a unique $a_b \in \mathcal{A}$ mapping $b_0 \longmapsto b$,*

(A.3) ***pluralistic:*** *there exist $a_1, a_2 \neq a_0$ such that $a_1^k \neq a_2^j$ for all positive integers k, j,*

(A.4) ***injective:*** *$a(b_1) \neq a(b_2)$ for all $a \in \mathcal{A}$ and $b_1 \neq b_2 \in \mathcal{B}$,*

(A.5) ***commutative:*** *$a_1 \circ a_2 = a_2 \circ a_1$ for all $a_1, a_2 \in \mathcal{A}$, and*

(A.6) ***acyclic:*** *$a^k \neq a$ for all $a \in \mathcal{A}$, $k > 1$.*

Discussion of Axioms. The properties enumerated in Theorem 1 consist of three regularity conditions (A.1-3) and three substantive axioms (A.4-6). First, countability (A.1) is imposed in parallel with the discreteness/continuity condition in definition of Bayesianism. Pluralism (A.3) is a richness condition which, as is made precise in Section 4, requires that an agent's beliefs can be pushed in multiple 'directions.' A simple example that fails (A.3) while meeting the other criteria is $\mathcal{B} = \mathbf{Z}_{\geq 0}$ and $\mathcal{A} = \{b \longmapsto b + k \mid k \in \mathbf{Z}_{\geq 0}\}$. Such a learning rule acts as an integer-valued counter that only increases; as with all cases that fail (A.3), any attempt to pair it with a Bayesian would be thwarted by the Bayes-plausibility condition that prior beliefs must lie in the convex hull of the corresponding set of posteriors.

The self-recording property (A.2) requires that there be some belief $b_0 \in \mathcal{B}$ from which all beliefs, including b_0, can be reached via a unique argument. Any such b_0 is termed an *ur-prior*[7] and models an 'original' or 'zero-information' state of knowledge. In a Bayesian learning rule, the prior distribution over the state space always serves this role. There is no requirement that b_0 be unique. For example, with $\mathcal{B} = \mathbf{Z}$ and $\mathcal{A} = \{b \longmapsto b + k \mid k \in \mathbf{Z}\}$, like for Agents 1 and 4 in the previous section, *every* belief satisfies the definition of an ur-prior. A practical implication of the self-recording property is that all algebraic structure described by the pair $(\mathcal{A}, \mathcal{B})$ is equally well encoded in the argument set $\mathcal{A}$ alone. Because each belief state b is associated with a unique argument a_b, the value of $a(b)$ for any arbitrary $a \in \mathcal{A}$ is equivalent to $a(a_b(b_0)) = (a_b \circ a)(b_0)$. The arguments do not need the beliefs to record their effect; they are 'self recording.' When comparing two self-recording learning rules, it suffices to focus solely on the compositional structure of their

[7] The 'ur-' terminology distinguishes it from the way in which any belief can be considered a 'prior' when it is the input to an argument. The prefix 'ur-' describes a primitive or original version of something, c.f. 'ur-text' etc.

arguments; as stated formally below, this consequence of (A.2) greatly simplifies the proof of Theorem 1.

Fact 1. *Self-recording learning rules* $(\mathcal{A}, \mathcal{B})$ *and* $(\mathcal{A}', \mathcal{B}')$ *are isomorphic if and only if* $\mathcal{A}$ *and* $\mathcal{A}'$ *are isomorphic as semigroups. Proof in the appendix.*

Although beyond the scope of this paper, departures from the self-recording property allow for more exotic structures. Intuitively, a learning rule that fails (A.2) either has an argument set strictly richer than its belief set, has a belief set strictly richer than its argument set, or both. For example, $\mathcal{B} = \mathbf{Z}$ is single dimensional while $\mathcal{A} = \{b \longmapsto c \cdot b + d \mid c, d \in \mathbf{Z}\}$ has two degrees of freedom; alternatively, $\mathcal{B} = \mathbf{Z} \times \mathbf{Z}$ is two dimensional while $\mathcal{A} = \{(b_1, b_2) \longmapsto (b_1 + k, b_2) \mid k \in \mathbf{Z}\}$ is single dimensional.

The substantive axioms of Theorem 1 are straightforward. Commutativity (A.5) demands that the semantic content of an argument is invariant to whatever other arguments have preceded it. Injectivity (A.4) and acyclicality (A.6) can be both understood as prohibitions against forms of memory loss. By (A.4), whenever an agent is known to have reached posterior belief b after receiving argument a, her prior belief is uniquely determined. By (A.6), whenever an agent is known to have transitioned from b_1 to b_2 after receiving multiple instances of argument a, the number of instances is uniquely determined.

3.1. Proof of Theorem 1

Summary. Given $(\mathcal{A}, \mathcal{B})$ satisfying (A.1-6), we progressively augment the structure of $\mathcal{A}$ to include inverses and fractional elements, enriching $\mathcal{A}$ from a semigroup to an Abelian group to a vector space over the rational numbers. We leverage the properties of vector spaces to construct an embedding of $\mathcal{A}$ into the real line in such a way that includes positive and negative numbers in its image. Finally, we interpret the image of $\mathcal{A}$ in $\mathbf{R}$ as log likelihood-ratios and identify a Bayesian learning rule whose associated joint probability distribution matches them. The proof is completed by verifying that Bayesian learning rules satisfy all six axioms.

Inverse and Fractional Arguments. First, let $(\mathcal{A}, \mathcal{B})$ be a learning rule satisfying (A.1-6). Because all Bayesians are self-recording (see end of proof), and by Lemma 1, it suffices to exhibit a Bayesian learning rule with an isomorphic argument set. We first insert new elements into $\mathcal{A}$ in order to guarantee the existence of a basis, which proves useful for manipulating the original set.

The initial challenge is to embed $\mathcal{A}$ into an Abelian (commutative) group which, unlike semigroups, must have an identity element and inverse elements. Note that properties (A.2) and (A.5) already guarantee an identity argument: as there is a unique $a^* \in \mathcal{A}$ mapping $b_0 \longmapsto b_0$, commutativity provides that for any b,

$$a^*(b) = a^*(a_b(b_0)) = a_b(a^*(b_0)) = a_b(b_0) = b,$$

so $a^* = a_{\mathsf{id}}$ is the identity argument. In service of establishing inverse elements, we first show that $\mathcal{A}$ satisfies a related form of injectivity: for any two beliefs b and b' there exists *at most one* argument mapping $b \longmapsto b'$.

Lemma 1. *If $\mathcal{A}$ is self-recording, injective and commutative, then $a_1(b) \neq a_2(b)$ for all $a_1 \neq a_2 \in \mathcal{A}$ and $b \in \mathcal{B}$. Proof in the appendix.*

The upshot of Lemma 1 is that $\mathcal{A}$ must be *cancellative*: the equation $a_1 \circ a_3 = a_2 \circ a_3$ always implies $a_1 = a_2$, even though a_3 may not have an inverse.[8] We can see this from the contrapositive: if $a_1 \neq a_2$, then $a_1(b) \neq a_2(b)$ by Lemma 1, and

$$(a_1 \circ a_3)(b) = a_3(a_1(b)) \neq a_3(a_2(b)) = (a_2 \circ a_3)(b)$$

by (A.4). The cancellation property allows us to apply a standard procedure in abstract algebra, the Grothendieck construction, to extend $\mathcal{A}$ to an Abelian group. At a high level, this procedure conceives of the pair (a_1, a_2) as representing 'a_1 *minus* a_2,' establishes an appropriate equivalence relation on pairs, and provides for an embedding of $\mathcal{A}$ into the set of equivalence classes $\mathcal{A}^+$, which is shown to be an Abelian group. Additionally, $\mathcal{A}^+$ consists only of the images of elements of $\mathcal{A}$ and their (missing) inverses. Hence $\mathcal{A}^+$ itself is both countable and acyclic.

Lemma 2. *If $\mathcal{A}$ is a countable, commutative, and cancellative semigroup, then there exists an injective homomorphism $f : \mathcal{A} \longrightarrow \mathcal{A}^+$, where $\mathcal{A}^+$ is a countable Abelian group. Moreover, any $a \in \mathcal{A}^+$ can be expressed $a = f(a_1) \circ (-f(a_2))$, where $a_1, a_2 \in \mathcal{A}$. Proof in the appendix.*

Now $\mathcal{A}^+$ is extended by filling in fractional elements. As a countable and acyclic Abelian group, $\mathcal{A}^+$ is isomorphic to countably many products of the set of integers. It therefore neatly

[8] Technically the cancellation property also requires that $a_3 \circ a_1 = a_3 \circ a_2$ implies $a_1 = a_2$, but for commutative semigroups this is an equivalent statement.

embeds inside a corresponding product of multiple copies of the rational numbers, denoted $\mathcal{A}^*$, which is a rational vector space. The dimension of $\mathcal{A}^*$ is unique and is denoted $d(\mathcal{A})$. Moreover, the embedding $\mathcal{A}^+ \longrightarrow \mathcal{A}^*$ is what I will term *essential*, in that the image of $\mathcal{A}^+$ in $\mathcal{A}^*$ contains a basis for $\mathcal{A}^*$. A small bit of additional work establishes an essential extension from $\mathcal{A}$ itself into $\mathcal{A}^*$.

Lemma 3. *If $\mathcal{A}$ is a countable, commutative, cancellative, and acyclic semigroup, then there exists an essential embedding $\mathcal{A} \longrightarrow \mathcal{A}^*$, where $\mathcal{A}^*$ is a rational vector space of countable dimension. Proof in the appendix.*

Embedding into the Reals. Now we establish an injective homomorphism $\mathcal{A}^* \longrightarrow \mathbf{R}$ via the following construction. Let $(\bar{a}_i)_{i=1}^{d(\mathcal{A})}$ be a basis of $\mathcal{A}^*$ contained in the image of $\mathcal{A}$ embedded in $\mathcal{A}*$. Every $a \in \mathcal{A}^*$ is thus of the form $a = \sum_{i=1}^{d(\mathcal{A})} q_i(a)\bar{a}_i$.[9] Note that by the pluralistic axiom (A.3), there must be at least two distinct basis elements.[10] We use the following process to construct a set of non-zero real numbers $(x_i)_{i=1}^{d(\mathcal{A})}$ that are 'mutually irrational': $x_i \notin \sum_{j\neq i} x_j\mathbf{Q}$, $x \neq 0$ for all $i = 1, \dots, d(\mathcal{A})$. First, set $x_1 = 1$. Choose $x_2 \notin \mathbf{Q}, x_2 < 0$, then choose any $x_3 \notin \mathbf{Q} + x_2\mathbf{Q}$, $x_r \notin \mathbf{Q} + x_2\mathbf{Q} + x_3\mathbf{Q}$, etc. The mapping

$$a = \sum_{i=1}^{d(\mathcal{A})} q_i(a)\bar{a}_i \longmapsto \sum_{i=1}^{d(\mathcal{A})} q_i(a)x_i$$

is an injective homomorphism. Composing the maps at each stage of the proof so far, we also have an injective homomorphism

$$\mathcal{A} \longrightarrow \mathcal{A}^+ \longrightarrow \mathcal{A}^* \longrightarrow \mathbf{R}.$$

As the image of $\mathcal{A}$ in $\mathcal{A}^*$ contains the basis $(\bar{a}_i)_{i=1}^{d(\mathcal{A})}$, and $\bar{a}_1$ is mapped to $x_1 > 0$ while $\bar{a}_2$ is mapped to $x_2 < 0$, the image of $\mathcal{A}$ in $\mathbf{R}$ necessarily contains both positive and negative numbers, so $\mathcal{A} \longrightarrow \mathbf{R}$ is termed *two-sided.*

The Log-Likelihood Connection. As a small detour, consider a *Bayesian* learning rule corresponding to $\{X, (Y_i)_{i=1}^{\infty}\}$, and let x_0 denote an arbitrarily chosen numeraire state. The log

[9] Note that $\{\bar{a}_i\}_{i=1}^{d(\mathcal{A})}$ is a *Hamel* basis of $\mathcal{A}^*$, which means that for any a all but finitely many of the $q_i(a)$ coefficients are zero. This eliminates the need to consider the cases of $d(\mathcal{A}) < \infty$ and $d(\mathcal{A}) = \infty$ separately.

[10] Otherwise, all arguments could be expressed $a = q(a)\bar{a}_1$, and hence for $q(a) = p(a)/r(a)$, $p(a), r(a) \in \mathbf{Z}_{>0}$, it would follow $a^{r(a)} = \bar{a}_1^{p(a)}$, contradicting (A.3).

odds-ratio of any state x relative to x_0 under belief π is

$$o(\pi)(x|x_0) = \log \frac{\pi[x]}{\pi[x_0]},$$

and the log likelihood-ratio of any state x relative to x_0 given signal sequence $y_1, \ldots, y_n$ is

$$l(y_1, \ldots, y_n)(x|x_0) = \sum_{i=1}^{k} \log \frac{\pi[y_i|x]}{\pi[y_i|x_0]}.$$

Applying these transformations to the beliefs and arguments, we obtain the logit transformation of Bayes' rule:

$$o(a_{(y_1,\ldots,y_n)}(\pi))(x|x_0) = o(\pi)(x|x_0) + l(y_1, \ldots, y_n)(x|x_0).$$

This rewriting of Bayes' rule demonstrates the one-to-one connection between Bayesian arguments and vectors of log likelihood-ratios. The application of any Bayesian argument can be seen as the addition of its corresponding log likelihood-ratio vector to the log-odds transformation of the prior.

It would be tempting to assume any learning rule which can be embedded as a subspace of $\mathbf{R}^n$ is therefore isomorphic to some Bayesian, but this ignores the Bayes-plausibility condition that the prior belief is constrained to lie in the convex hull of the posterior distributions. In general, 0 lying in the convex hull of a set of vectors in $\mathbf{R}^n$ does *not* imply that the image of 0 under a reverse logit transformation lies in the convex hull of the image of the vectors. However, this connection does hold for $n = 1$. Thus, any learning rule which can be embedded into $\mathbf{R}^1$ and whose range includes both positive and negative values is isomorphic to a Bayesian.

Lemma 4. *If $(\mathcal{A}, \mathcal{B})$ is a countable, self-recording learning rule and there exists an injective, two-sided homomorphism $\mathcal{A} \longrightarrow \mathbf{R}$, then $(\mathcal{A}, \mathcal{B})$ is a Virtual Bayesian. Proof in the appendix.*

To conclude: by Lemmas 1 – 2 there exists a two sided injective homomorphism $\mathcal{A} \longrightarrow \mathbf{R}$ and by Lemma 4 it follows $(\mathcal{A}, \mathcal{B})$ is a Virtual Bayesian. The other direction of the proof is straightforward and does not merit any special discussion.

Lemma 5. *All Bayesian learning rules satisfy axioms (A.1-6). Proof in the appendix.*

3.2. **Examples of Virtual Bayesians**

Virtual Bayesian learning rules are found widely among updating procedures analyzed by prior literature as well as among colloquial rule-of-thumb heuristics. As the following cases illustrate, Virtual Bayesians are often identified by updating in the form of tallying one or more quantities, shifting along a spectrum, or a combination of the two.

Probability Weighting and Variations. The most familiar instances of Virtual Bayesians are produced by coupling Bayes' rule with a given belief-distortion function. For instance, consider the case of probability weighting as used in Cumulative Prospect Theory. According to the functional form introduced in Prelec (1998), objective probabilities $p \in (0, 1)$ are transformed into weighted versions via $h(p) = e^{-(-\log p)^{\alpha}}$, where $\alpha \in (0, 1)$.

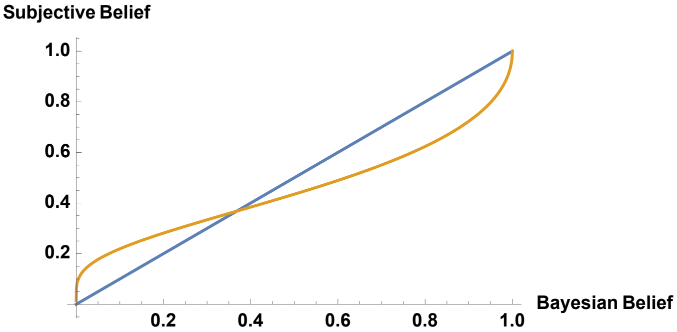

Figure 1. Probability Weighting Function. *The orange line describes an agent's subjective probability as a function of the objective value, where $\alpha = 1/2$. The 45-degree line is in blue.*

Any Bayesian learning argument can be warped into a Virtual Bayesian by application of the probability weighting function. For a given log likelihood-ratio l, where the Bayesian would update $p \longmapsto e^l/(1 + e^l)$, the probability-weighting version would update according to

$$\tilde{p} \longmapsto \exp\left\{-\left(-\log\left(\frac{e^l}{e^{(-\log\tilde{p})/\alpha} - (1 - e^l)}\right)\right)^{\alpha}\right\}.$$

The key point is that expressing miscalibrated probability assessments – overstating low events and understating high ones – is compatible with the kind of internally consistent updating exhibited by bona fide Bayesians.

Theorem 1 also enables us to identify Virtual Bayesians merely from their functional forms, which can then point the way to underlying belief transformation functions. Consider the learning rule where $\mathcal{A} = \{b \longmapsto b^x \mid x \in \mathbf{Q}_{>0}\}$, and $\mathcal{B} = \{a(1/2) \mid a \in \mathcal{A}\}$. As is readily verified, this satisfies the axioms of Theorem 1, so it is a Virtual Bayesian, and the transformation from the Bayesian belief set to the power-law updater is easily obtained: $b_{\textsf{power-law}} = 2^{1-1/b_{\textsf{Bayesian}}}$. Figure 2 illustrates how it looks qualitatively like the opposite of a probability weighting function.

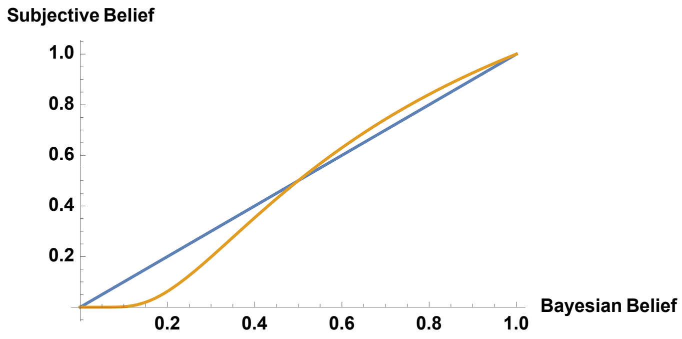

Figure 2. Power-Law Belief Distortion. *The orange line describes an agent's subjective probability as a function of the objective value, where $\alpha = 1/2$. The 45-degree line is in blue.*

Reinforcement Learning. The learning-in-games literature features several heuristics by which players transition among strategies according to the results of past play. One of the earliest studied is reinforcement learning, by which one maintains a weight value for each possible action, chooses new actions with probabilities determined by the weights, and updates the weights according to realized payoffs. As formalized in Erev and Roth (1998), action α_k is initially given weight $q_k = 1$; thereafter, whenever action α_k is selected and the player receives payoff x, he updates k's weight by $q_k \longmapsto q_k + R(x)$, where $R(\cdot) \geq 0$ is an increasing 'reward' function. Being defined in the context of games, reinforcement learning also specifies a particular action function (α_k is given probability $q_k / \sum_j q_j$), but, given the present paper's focus on learning rules, we can consider its updating procedure in isolation.

A reinforcement-learning agent has beliefs in the form $b = (q_1, \ldots, q_n)$, and arguments describe what happens to the weights after potential play of the game:

$$a : (q_1, \ldots, q_n) \longmapsto \left(q_1 + \sum_{j=1}^{m_1} R_{1,j}, \ldots q_n + \sum_{j=1}^{m_n} R_{n,j} \right),$$

where each reward $R_{k,j}$ is equal to $R(u(\alpha_k, \alpha_{\mathsf{opp}}))$ for some opponent action profile α_{opp}. $\mathcal{A}$ is the set of all such arguments, and $\mathcal{B} = \mathcal{A}(b_0)$ for $b_0 = (1, 1, \ldots, 1)$.

The axioms of Theorem 1 are easily verified, verifying that the learning rule generated by reinforcement learning is Virtual Bayesian. Intuitively, beliefs take the form of a multi-dimensional tally, one of the paradigms of Virtual Bayesians. What is perhaps unintuitive about this example is that a reinforcement-learning agent is explicitly not calculating probabilities, and the player's weights and updates have nothing to do with frequency predictions. Nonetheless, the *compositional* structure created by the heuristic is entirely Bayesian.

Fictitious Play. A distant cousin of reinforcement learning is fictitious play, by which a player in a game best-responds to the empirical frequency of her opponents' actions. Although players keep track of *opponent* actions rather than *own* actions, the updating in fictitious play also forms a multi-dimensional tally. That is, each argument specifies

$$a : (\alpha_{\mathsf{opp},1}, \ldots, \alpha_{\mathsf{opp},n}) \longmapsto (\alpha_{\mathsf{opp},1} + \Delta_1, \ldots, \alpha_{\mathsf{opp},n} + \Delta_n),$$

where Δ_k is the number of times the player observes profile $\alpha_{\mathsf{opp},k}$ in a given temporal window.

Pro/Con Lists. As demonstrated by the example of the woman reasoning about flood risk in the Introduction and Agent 4 in Section 2, Virtual Bayesians need not have beliefs in the form of probability distributions. Indeed, many commonplace procedures for keeping track of information use natural language, symbols or other structures which lack a clear probabilistic interpretation. Consider pro-con lists. An agent wishing to evaluate the quality of a good maintains two columns of features about the object, one recording its virtues and the other its faults. The agent's belief pairs the total number of pro items and the total number of con items. Formally, $\mathcal{B} = \{(p, c) \mid p, c \geq 0\}$ and $\mathcal{A} = \{(p, c) \longmapsto (p + p_{\mathsf{new}}, c + c_{\mathsf{new}}) \mid p_{\mathsf{new}}, c_{\mathsf{new}} \geq 0\}$. This learning rule is isomorphic to that of Agent 2. More generally, any evaluation rubric in which a number is assigned to each of a finite number of categories and arguments take the form of adjustments to each category satisfies axioms (A.1-6).

3.3. **Examples of Non Virtual Bayesians**

Learning rules which cannot be paired with an isomorphic Bayesian are similarly ubiquitous among the updating procedures in the literature and everyday life. The examples below demonstrate failures of the three substantive axioms of Theorem 1. We examine failures of commutativity first, followed by non-injective and then non-acyclic learning rules.

DeGroot Learning. An intuitive procedure for incorporating new information into one's current beliefs is averaging. Although DeGroot (1974) considered averaging only with respect to a fixed network of agents, it seems appropriate to label any case of updating via averaging as a form of 'DeGroot' learning. Formally, let an agent's set of beliefs be $\mathbf{Q}$, and, for a fixed value of $\alpha \in (0, 1)$, let $\mathcal{A}$ be the set of all maps

$$a_{q'} : q \longmapsto \alpha q' + (1-\alpha)q, \quad q' \in \mathbf{Q}.$$

That is, the DeGroot agent incorporates a new opinion q' into his current belief q by taking a weighted average of the two with weight α on the new opinion. Although the agent's beliefs comprise a single-dimensional spectrum, as in several examples of Virtual Bayesians encountered thus far, he does not transition among the beliefs by shifting but rather by compressing the spectrum towards a particular point (the new opinion). This creates order effects, which violate commutativity. For example, for $\alpha = 1/2$,

$$a_{q_1'} \circ a_{q_2'} : q \longmapsto \frac{q + q_1' + 2q_2'}{4} \quad \text{but} \quad a_{q_2'} \circ a_{q_1'} : q \longmapsto \frac{q + 2q_1' + q_2'}{4}.$$

The root of the non-Bayesian character lies in the fixed weighting. If we were to amend the procedure by allowing for suitably adjustable weights, the heuristic would become Virtual Bayesian. Specifically, set $\mathcal{B} = \mathbf{Q} \times \mathbf{N}$, where each belief couples an opinion q with a strength n. Each argument is identified by an additional (q', n'), and updating is of the form

$$(q, n) \longmapsto \left(\frac{nq + n'q'}{n + n'}, n + n' \right).$$

The flexible weighting means that each argument's associated opinion is factored into the agent's ultimate posterior according to its associated strength, and not according to the order in which it was received.

Other Cases of Non-Commutative Learning Rules. An agent engaging in base-rate neglect (Benjamin, Bodoh-Creed and Rabin, 2019) has beliefs in the form of probabilities, but updates with a modified version of Bayes' rule that underweights her prior belief. Specifically, for $\pi \in \mathcal{B}$, any state $x \in \mathcal{X}$, and any signal $y \in \mathcal{Y}$,

$$\pi(x) \longmapsto \frac{\pi(x)^\alpha \cdot \mathbf{P}[\hat{y}|x]}{\sum_{x' \in \mathcal{X}} \pi(x')^\alpha \cdot \mathbf{P}[\hat{y}|x']},$$

where $\alpha < 1$ captures the extent of underweighting the base-rate. As with DeGroot learning, this differential treatment of one's prior relative to the likelihood creates order effects and violates commutativity. Examples are also found in the learning-in-games literature. Consider the *experience-weighted attraction* (EWA) model of Camerer and Ho (1999), which nests reinforcement learning and fictitious play under a single framework. Unlike its two special cases, the EWA model features a discount rate φ on the weight a player assigns to any given strategy, which 'depreciates previous attraction' to that strategy. This and another depreciation rate generically lead to order effects.

As a final example, consider *concatenating* learning rules, where, for a set of symbols S, beliefs are sequences $(s_1, \ldots, s_n),\ s_i \in S$ and arguments correspond to additional sequences that the agent appends to her current belief:

$$\mathcal{A} = \{(s_1, \ldots, s_n) \longmapsto (s_1, \ldots, s_n, \hat{s}_1, \ldots, \hat{s}_m) \mid \hat{s}_k \in X\}\,.$$

A concatenating learning rule satisfies all of the axioms of Theorem 1 except for commutativity (A.5). Note that unlike the other examples, which feature heuristics or biases, concatenation records *strictly more* information than a Bayesian does.

Deductive Reasoning. We next consider failures of injectivity, which necessarily involve 'collapsing' multiple priors onto a single posterior. One common example is *deductive* reasoning, whereby an agent updates about an unobserved state by progressively ruling out different possibile state values. Each belief is a subset of already-disproven states, and each argument corresponds to a subset of newly-disproven states. As the agent learns, she narrows the range of possibilities until she reaches either a single state or concludes that none is possible.

Definition. *Learning rule $(\mathcal{A}, \mathcal{B})$ is* ***deductive*** *if there exists a state space $\mathcal{X}$ such that (1) each belief is a subset of disproved states $b \subset \mathcal{X}$ and (2) each argument is associated with a subset of additionally impossible states: $a : b \longmapsto b \cup X_a$ for some $X_a \subset \mathcal{X}$.*

Deductive reasoning shares a close relationship with Bayesian updating, but they are fundamentally different. The similarity is that a Bayesian who receives signals with *zero* likelihood values for some states appears to rule out those states just as a deductive reasoner would. However, Bayes' rule provides no guidance for how to update following a signal that the agent believes has zero probability. The deductive reasoner, in an equivalent circumstance, simply transitions to believe that no state is possible. Furthermore, as shown in the following Proposition, deductive reasoners have a very different compositional structure. In particular, whereas a Bayesian distinguishes between different multiples of a given argument, a deductive reasoner treats all repeated instances of an argument as equivalent to a single instance. Her learning rule is ***idempotent***: $a \circ a = a$ for all $a \in \mathcal{A}$.

Proposition 1. *A self-recording learning rule is isomorphic to some deductive learning rule if and only if it is commutative and idempotent. Proof in the appendix.*

Though dense, the proof of Proposition 1 admits an intuitive interpretation. To show that an arbitrary learning rule is deductive, one needs to exhibit a set of states $\mathcal{X}$ such that each argument corresponds to a subset of states. As a thought experiment, suppose the task were already completed. Then, for every element $x \in \mathcal{X}$, there would be a set $A_x \subset \mathcal{A}$ of *arguments* whose corresponding subsets contained x. Any collection of arguments generated in this way would have the following inclusion property: whenever $a \in A_x$ and $a \circ a' \circ a'' = a' \circ a''$, at least one of a' or a'' must be contained in A_x. Intuitively, this says that whenever (1) argument a 'communicates' state x and (2) arguments a' and a'' together subsume the information contained in a, at least one of the subsuming arguments a' or a'' must themselves communicate state x. Any underlying state x is therefore functionally equivalent to a subset of arguments A_x with the inclusion property. The proof of Proposition 1 demonstrates that we can define $\mathcal{X}$ to be the set of *all* argument subsets with the inclusion property.

As a final note, one might suppose that the structure of a deductive learning rule is sufficient to recover the underlying state space (up to relabeling), but a simple counterexample shows otherwise. Consider $X = \{x, y\}$ and $\mathcal{A}$ consisting of a_0 (corresponding to $\emptyset$), a_1 ($\{x\}$), a_2 ($\{y\}$), and a_3 ($\{x, y\}$). Compare this with $X' = \{x, y, z\}$ and $\mathcal{A}'$ comprising a'_0 ($\emptyset$), a'_1 ($\{x, z\}$), a'_2 ($\{y, z\}$), and a_3 ($\{x, y, z\}$). The learning rules generated by these two argument sets are isomorphic, and hence there is no way to identify from the compositional structure of arguments alone whether the state space is X or X'. In both cases, the compositional structure is described by $a_1 \circ a_2 = a_3$ (or $a'_1 \circ a'_2 = a'_3$), but while this equation reveals that neither of a_1 or a_2's associated state sets is a subset of the other, there is nothing to indicate whether their intersection is nonempty.

Adopting the Strongest Opinion. Another variant on DeGroot learning specifies $\mathcal{B} = \mathbf{Q} \times \mathbf{N}$, with beliefs consisting of opinion-strength pairs as in the adjustable-weight example, but has arguments of the form

$$(q, n) \longmapsto \begin{cases} (q', n') \text{ if } n' > n \\ (q, n) \text{ otherwise.} \end{cases}$$

That is, the agent keeps track of the relative strength of any opinion, but instead of giving more weight to a stronger new opinion, the agent either adopts it fully if it has greater strength than one's current opinion or discards it otherwise. Like with deductive reasoning, all arguments are idempotent, violating injectivity. Furthermore, this variant also violates commutativity, but such violations are restricted to corner cases in which multiple arguments share identical opinion strengths.

Cyclic Learning Rules. Consider an agent who learns about the status of a light bulb. The set of beliefs is $\mathcal{B} = \{\mathsf{on}, \mathsf{off}\}$, and the set of arguments consists of

$$\mathsf{maintain} : \begin{cases} \mathsf{on} \longmapsto \mathsf{on} \\ \\ \mathsf{off} \longmapsto \mathsf{off} \end{cases} \quad \text{and} \quad \mathsf{switch} : \begin{cases} \mathsf{on} \longmapsto \mathsf{off} \\ \\ \mathsf{off} \longmapsto \mathsf{on}. \end{cases}$$

More generally there are 'cyclic' learning rules with $\mathcal{B} = \{\mathsf{state}_0, \ldots, \mathsf{state}_{n-1}\}$ and

$$\mathcal{A} = \{\mathsf{state}_k \longmapsto \mathsf{state}_{k+j \text{ modulo } n} \mid j = 0, \ldots, n-1\}\,.$$

These model uncertainty in environments with an inherently cyclic structure, such as those dealing with recurring blocks of time. Their argument semigroups are isomorphic to $\mathbf{Z}_n = \mathbf{Z}/n\mathbf{Z}$. In fact there is a close connection between Bayesian and cyclic learning rules. While no finite learning rule can be acyclic, and therefore no finite learning rule can be a Virtual Bayesian, it is still possible for a learning rule to satisfy the other five axioms of Theorem 1. Those who do have an fundamentally cyclic structure.

Proposition 2. *A finite, self-recording learning rule* $(\mathcal{A}, \mathcal{B})$ *is pluralistic (A.3), injective (A.4), and commutative (A.5) if and only if there exist primes* $p_1, \ldots, p_k$*,* $k \geq 2$*, such that* $\mathcal{A} \cong \mathbf{Z}_{p_1} \times \cdots \times \mathbf{Z}_{p_k}$*.* *Proof in the appendix.*

4. Geometric Properties of Virtual Bayesians

The paper has thus far considered purely algebraic relationships between arguments. For example, if $a_2 = a_1 \circ a_1$, then a_2 is understood as 'twice' a_1 and a_1 is 'half' a_2, etc. But what can be said more generally, especially for arguments that lack a clear algebraic connection? This section explores one particular relationship between two arguments: *agreement*, the extent to which each supports the effect of the other. It is shown that a learning rule is Virtual Bayesian if and only if (subject to regularity axioms) it admits an agreement function that is *additive.* Moreover, an additive agrement function serves as the foundation for embedding the set of arguments into Euclidean space. Hence, Virtual Bayesian arguments are characterized by various geometric measures; this section illustrates how those measures complement more familiar probabilistic concepts.

4.1. **Additive Agreement**

What does it mean for two arguments to 'agree' with each other? That is, what should be the function form of $\gamma : \mathcal{A} \times \mathcal{A} \longrightarrow \mathbf{R}$, where $\gamma(a_1, a_2)$ is the level of agreement $(\gamma(\cdot,\cdot) > 0)$ or disagreement $(\gamma(\cdot,\cdot) < 0)$ between a_1 and a_2? Consider the following thought experiment. An agent with learning rule $(\mathcal{A}, \mathcal{B})$ is presented with arguments a_1 and a_2, which have agreement $\gamma(a_1, a_2)$. Next, suppose a_1 were decomposed into components a_1' and a_1'' such that $a_1 = a_1' \circ a_1''$. There would be agreement $\gamma(a_1', a_2)$ between a_1' and a_2 as well as $\gamma(a_1'', a_2)$ between a_1'' and a_2. Intuitively, what agreement a_1 originally had with a_2 stemmed in some part from the a_1' and in some part from a_1''. It follows that $\gamma(a_1', a_2)$ and $\gamma(a_1'', a_2)$ ought to sum to $\gamma(a_1, a_2)$.

Say that $\gamma : \mathcal{A} \times \mathcal{A} \longrightarrow \mathbf{R}$ is an *additive agreement function* if it satisfies the above decomposition insensitivity – $\gamma(a_1', a_2) + \gamma(a_1'', a_2) = \gamma(a_1' \circ a_1'', a_2)$ for all $a_1', a_1'', a_2 \in \mathcal{A}$ – as well as two regularity conditions. First, γ is symmetric: $\gamma(a_1, a_2) = \gamma(a_2, a_1)$ for all $a_1, a_2 \in \mathcal{A}$. Second, it is *positive definite*:

$$\gamma(a_1, a_2) < \frac{1}{2}\left(\gamma(a_1, a_1) + \gamma(a_2, a_2)\right)$$

for all $a_1, a_2 \neq a_{\mathsf{id}}$. If $\mathcal{A}$ contains inverses, then positive definiteness is equivalent to requiring that all arguments agree with themselves: $\gamma(a, a) > 0$ for $a \neq a_{\mathsf{id}}$.[11] If $\mathcal{A}$ does not contain inverses, the positive definiteness not only implies all arguments have positive self agreement, but it also ensures that, were $\mathcal{A}$ to be augmented by the insertion of inverse arguments, the extension of γ would still assign all arguments positive self agreement.

Additive agreement functions provide a link between Virtual Bayesians and embeddings into $\mathbf{R}^n$. The key factor is that the three defining properties of additive agreement functions are also satisfied by an inner product on a real vector space. An argument set $\mathcal{A}$ is not a rich enough space on which to define a true inner product, but an additive agreement function provides the foundation for extending $\mathcal{A}$ into $\mathbf{R}^n$.

[11] One implication of the additivity condition is that $\gamma(a_{\mathsf{id}}, a_{\mathsf{id}}) = 0$, so this case must be exempted.

Additive Agreement and Virtual Bayesians. Being able to define an additive agreement function on a given learning rule $(\mathcal{A}, \mathcal{B})$ guarantees that $\mathcal{A}$ satisfies axioms (A.4,5,6) from Theorem 1. Commutativity (A.5) and acyclicality (A.6) follow from the fact that the real numbers, as a group, share these properties. Similarly, a learning rule with an additive agreement function necessarily satisfies the cancellation property, which is shown to imply injectivity (A.4). Given any Virtual Bayesian learning rule, it is straightforward to define an additive agreement function, e.g. by treating its basis as an orthonormal set. Thus the existence of an additive agreement function, in conjunction with axioms (A.1,2,3), completely characterizes Virtual Bayesians.

Theorem 2. *Learning rule $(\mathcal{A}, \mathcal{B})$ is a Virtual Bayesian if and only if it is (A.1) countable, (A.2) self-recording, (A.3) pluralistic, and it admits an additive agreement function $\gamma : \mathcal{A} \times \mathcal{A} \longrightarrow \mathbf{R}$. Proof in the appendix.*

Additive Agreement and Euclidean Embeddings. Any embedding $f : \mathcal{A} \longrightarrow \mathbf{R}^n$ defines an agreement function by pulling back the standard inner product on $\mathbf{R}^n$: $\gamma(a_1, a_2) \equiv \langle f(a_1), f(a_2) \rangle$. As is shown by Theorem 3 below, all additive agreement functions can be generated in this way. The proof of Theorem 2 establishes that γ on $\mathcal{A}$ extends uniquely to the rational completion $\mathcal{A}^*$. There are multiple ways to embed $\mathcal{A}^* \longrightarrow \mathbf{R}$, but only a subset of them preserve γ; one such way is the inclusion map $\mathcal{A}^* \longrightarrow \mathbf{R}^n$. However, it is also possible to embed $\mathcal{A}^*$ into any lower dimensional real vector space by collapsing distinct dimensions of $\mathcal{A}^*$ onto a single dimension of $\mathbf{R}^n$. For example, let $(\mathcal{A}, \mathcal{B})$ be a self-recording learning rule with $\mathcal{A} \cong \mathbf{Z}_{\geq 0} \times \mathbf{Z}_{\geq 0}$. That is, from an algebraic perspective, $\mathcal{A}$ is generated from two distinct elements. As illustrated by Agent 2, $(\mathcal{A}, \mathcal{B})$ could represent a Beta-Binomial learning model in which each of the generating arguments, H and T, each communicate information that does not directly contradict the other (in the Beta-Binomial model, both H and T signals lower the learner's subjective variance about the state). Under this interpretation, the most intuitive specification of an agreement function sets $\gamma(\mathsf{H}, \mathsf{T}) = 0$, corresponding to an embedding $\mathcal{A} \longrightarrow \mathbf{R}^2$. However, $(\mathcal{A}, \mathcal{B})$ could also represent uncertainty about a binary state, as with Agent 1. For example, if one of the generating arguments is H and the other is $\sqrt{2}\mathsf{T}$, then the natural specification of agreement sets $\gamma(\mathsf{H}, \sqrt{2}\mathsf{T}) = -\sqrt{2}$, corresponding to an embedding $\mathcal{A} \longrightarrow \mathbf{R}^1$. The multiple plausible interpretations in this example illustrate how the algebraic structure of $(\mathcal{A}, \mathcal{B})$ is made more concrete by specifying an additive agreement function.

Theorem 3. *Let $(\mathcal{A}, \mathcal{B})$ be a Virtual Bayesian learning rule. Then*

(a) *For every additive agreement function γ on $\mathcal{A}$ there is a unique $n \leq d(\mathcal{A})$ and a unique (up to orthogonal transformation) essential embedding $f : \mathcal{A} \longrightarrow \mathbf{R}^n$ extending γ to the standard inner product on $\mathbf{R}^n$.*

(b) *For every essential embedding $f : \mathcal{A} \longrightarrow \mathbf{R}^n$ there is a unique additive agreement function γ on $\mathcal{A}$ that f extends to the standard inner product on $\mathbf{R}^n$.*

(c) *There exists an essential embedding $\mathcal{A} \longrightarrow \mathbf{R}^n$ for all $1 \leq n \leq d(\mathcal{A})$.*

Proof in the appendix.

Statements (a) and (b) of Theorem 3 establish that there is a near one-to-one connection between additive agreement functions on $\mathcal{A}$ and embeddings into $\mathbf{R}^n$. The only exception is highly technical: starting with a given embedding $\mathcal{A} \longrightarrow \mathbf{R}^n$, one can flip, rotate, or otherwise rigidly transform the image of $\mathcal{A}$ inside $\mathbf{R}^n$ to manufacture a distinct embedding, but this is tantamount to imposing an alternative coordinate system. Similarly, by embedding $\mathcal{A}$ into $\mathbf{R}^n$ and then embedding $\mathbf{R}^n$ into $\mathbf{R}^{n+1}$, one obtains a technically distinct mapping $\mathcal{A} \longrightarrow \mathbf{R}^{n+1}$, but it is necessarily *not* an essential embedding. (Recall that $f : \mathcal{A} \longrightarrow \mathbf{R}^n$ is essential if $f(\mathcal{A})$ contains a basis for $\mathbf{R}^n$.) Finally, statement (c) confirms that $\mathcal{A}$ can be embedded into Euclidean space of any dimension n up to its 'algebraic dimension' $d(\mathcal{A})$. The reasoning is intuitive: by treating all basis elements of $\mathcal{A}$ as mutually orthogonal, one constructs an embedding $\mathcal{A} \longrightarrow \mathbf{R}^{d(\mathcal{A})}$; by treating more and more basis elements as linearly dependent, one can reduce the dimensionality arbitrarily to $n = 1$.

4.2. **Agreement Geometry**

When into $\mathbf{R}^n$, $\mathcal{A}$ obtains a handful of familiar geometric concepts. Interpreted as vectors in $\mathbf{R}^n$, arguments are equipped with the standard inner product; they have length, angle, and direction; one can specify the distance between two arguments or the projection of one onto another. Theorem 3 shows that the universe of different Euclidean geometries available to an learning rule are enumerated by the set of additive agreement functions.

The most primitive structure inherited by an learning rule is real-valued scaling: given argument $a \in \mathcal{A}$, any $k \in \mathbf{R}$ identifies a unique scaled version ka. For positive integer values of k, the scaled copy ka is part of $\mathcal{A}$ itself; for $k \in \mathbf{Q}$, ka lies in $\mathcal{A}^*$; for irrational values of k, ka is guaranteed *not* to be a part of $\mathcal{A}$, but it can nonetheless be understood as part of the linear extension of $\mathcal{A}$.

Scaling also defines a notion of argument magnitude via the norm $|a| \equiv \sqrt{\gamma(a,a)}$. This coincides with scaling, in that $|ka| = |k| \cdot |a|$.

Each argument is identified by the conjunction of its magnitude and direction; the direction of any argument $a \in \mathcal{A}$ is its counterpart $a/|a|$ lying on the unit sphere S^{n-1}. The angle between two arguments is

$$\theta(a_1, a_2) \equiv \cos\left(\frac{\gamma(a_1, a_2)}{|a_1| \cdot |a_2|}\right),$$

and the projection of one argument onto another is

$$a_1|a_2 \equiv \left(\frac{\gamma(a_1, a_2)}{|a_2|}\right) a_2.$$

Finally, distance between two arguments a_1 and a_2 is defined by the metric $|a_1 - a_2|$.

4.3. **Bayesian Comparison: Examples**

We can compare the various Euclidean structures that come with an additive agreement function to probabilistic notions already defined for a Bayesian learning rule by first examining several examples.

Single Binary Issue. Consider learning as modeled by Agent 1 in Section 2 (the 'Bernoulli Bayesian'), who has uncertainty about a binary state. Each argument $a \in \mathcal{A}$ is associated with a single log likelihood-ratio l_a. Every additive agreement function which embeds $\mathcal{A}$ into $\mathbf{R}$ is of the form $\gamma(a_1, a_2) = w \cdot l_{a_1} \cdot l_{a_2}$, where $w > 0$. To understand the relationship between arguments and probabilistic notions in this particular case, consider the effect of a given argument on a fixed reference prior. For example, if the reference prior is either $\mathbf{P}[\omega = 1] = 2/3$ or $\mathbf{P}[\omega = 1] = 1/3$, then the posterior as a function of the argument is described by the graph in Figure 3. There are two directions, positive and negative; scaling up any positive argument leads to a posterior belief arbitrarily close to $\mathbf{P}[\omega = 1] = 1$, and scaling up a negative argument leads the learner towards $\mathbf{P}[\omega = 1] = 0$. In this way the two directions of arguments are associated with the two distinct Bayesian states. Moreover, sufficiently extreme magnitude is associated with certainty on a particular state.

The relationship between intermediate scaling of an argument and the corresponding posterior distribution depends on which reference prior is chosen. As illustrated by Figure 3, scaling up a positive argument always increases the expectation of the probabilistic belief, and scaling up a

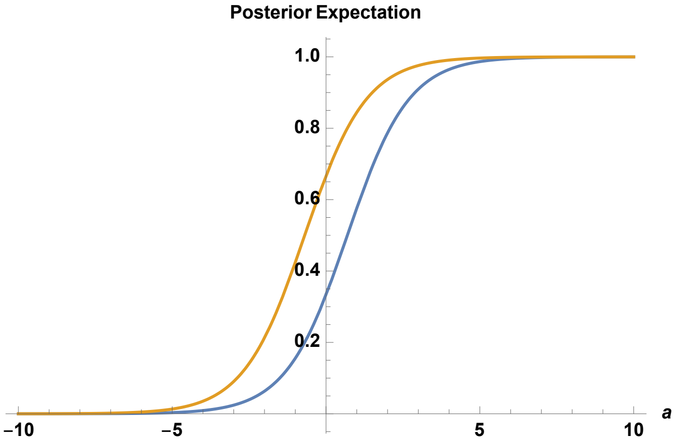

FIGURE 3. Posterior Expectation as a Function of Argument. *The orange line describes an agent's posterior probability given a prior of* $\mathbf{P}[\omega = 1] = 2/3$ *as a function of the argument* a *as embedded in* $\mathbf{R}$*. The blue line is for* $\mathbf{P}[\omega = 1] = 1/3$*.*

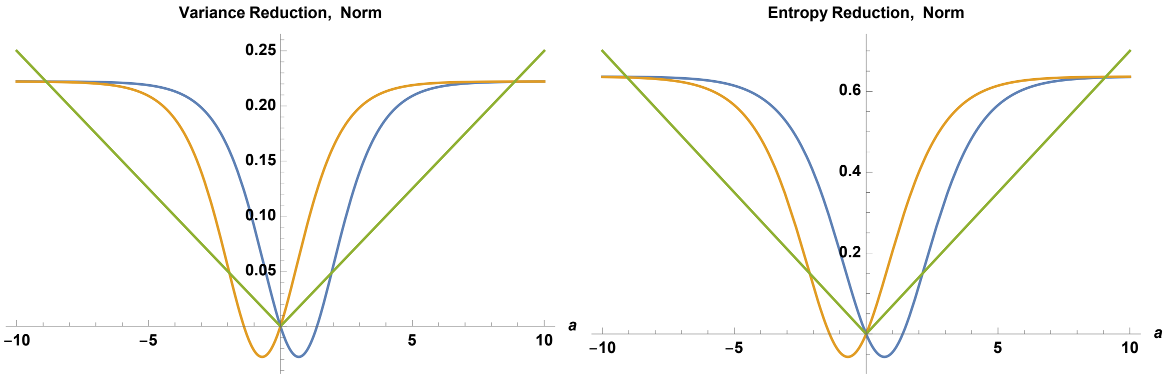

FIGURE 4. Variance, Entropy Reduction as a Function of Argument. *The left graph compares the reduction in variance for an agent with prior belief* $\mathbf{P}[\omega = 1] = 2/3$ *(orange line) or* $\mathbf{P}[\omega = 1] = 1/3$ *(blue line) as a function of argument. The green line describes the norm of the argument. The right graph shows corresponding relationships for entropy reduction.*

negative argument always decreases the expectation. The effect on variance and entropy, however, depends on whether the prior is above or below the uniform belief of $\mathbf{P}[\omega = 1] = 1/2$. Intuitively, measuring information gain by reduction in variance or entropy tracks how the posterior moves towards or away from certainty, while measuring information gain by the agreement norm tracks movement away from one's prior, which must always be positive. As illustrated in Figure 4, when the agent's prior is – for example – 1/3, small rightward arguments push the posterior closer to 1/2, and therefore have *negative* variance and entropy reductions.

Agreement itself, as specified by $\gamma(a_1, a_2) = w \cdot l_{a_1} \cdot l_{a_2}$, measures the extent to which two arguments push the prior belief in the same direction. Agreement is positive if and only if both arguments share the same sign – that is, if they both push the agent towards believing $\omega = 1$ or both push towards $\omega = 0$. The magnitude of agreement or disagreement is simply proportional to the product of the arguments' associated log likelihood-ratios.

Multiple Independent Binary Issues. Consider an extension of the two-state Bayesian learning model in which there are n binary issues, where beliefs are independent across issues and each argument preserves independence. Moreover assume that the arguments can lead to arbitrarily high certainty on any of the 2^n states. For this example there is a natural class of additive agreement functions which treat separate issues as orthogonal: set $\gamma(a_1, a_2) = \sum_{i=1}^{n} w^i l^i_{a_1} l^i_{a_2}$, where $w^i > 0$ are importance weights on the different issues. In the case of $n = 2$ such an embedding can be visualized in the following way, as illustrated in Figure 5. Take any argument embedded in $\mathbf{R}^2$, and consider what is the effect of the argument on a given reference prior, e.g. the uniform distribution.[12] This produces a mapping from $\mathbf{R}^2$ into the set of possible beliefs as represented by the unit square $\Delta(\{0,1\}) \times \Delta(\{0,1\})$.

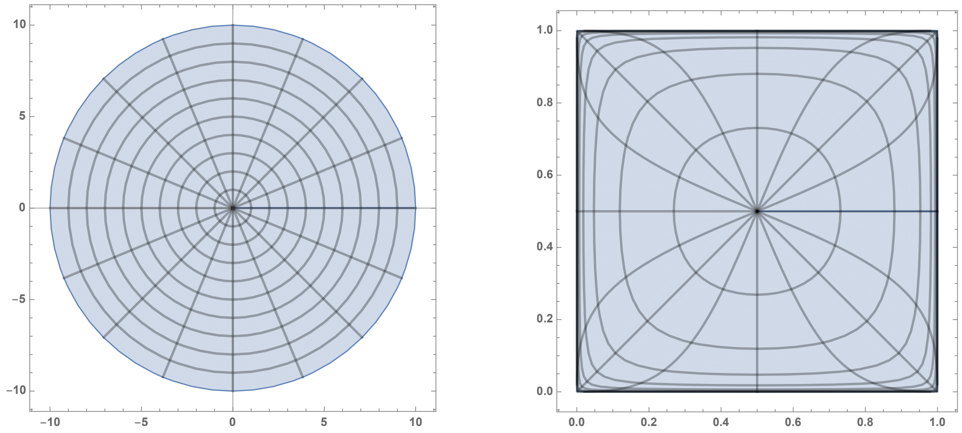

Figure 5. Standard Embedding of $\mathcal{A}$ into $\mathbf{R}^2$ for Two Independent Binary Issues. *The left figure shows polar contour lines in $\mathbf{R}^2$; the right figure shows the same contour lines under the following transformation: each argument as embedded in $\mathbf{R}^2$ is identified with the posterior distribution $p_{post} \in (0,1) \times (0,1)$ it produces when the prior is the uniform distribution.*

In contrast to the two-state example, there is a continuum of directions in $\mathbf{R}^2$. As before, almost all directions correspond to a unique limit belief, but now there are measure-zero borders separating $\mathbf{R}^2$ into four regions corresponding to the four different configurations of the two binary issues. Also

[12] The identity of the reference prior does not change the qualitative appearance of the graph, but using the uniform distribution as a reference makes the relationships clearer.

in this case, projection has a very intuitive meaning. Projecting an argument onto the axes isolates the information an argument communicates about each issue. Projecting onto an arbitrary direction forces the original argument to discuss the two issues at the relative frequency of the projection direction. As a way of illustrating how different embeddings treat the set of arguments differently, consider stretching one issue/dimension, e.g. by specifying that the argument originally mapped to $(1, 0)$ is instead mapped to $(2, 0)$, as in Figure 6. For arguments whose angle is sufficiently close to the stretched issue, angles between them become tighter with the stretching; for those closer to the other issue, the angles are widened.

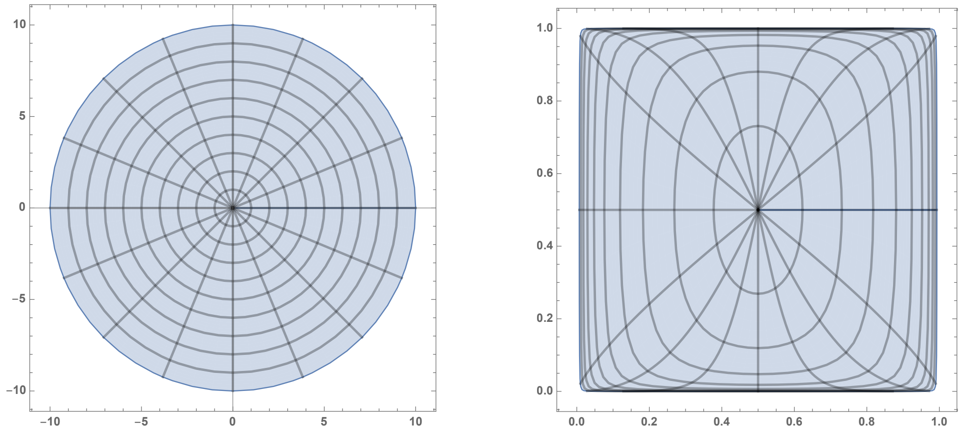

Figure 6. Stretched Embedding of $\mathcal{A}$ into $\mathbf{R}^2$ for Two Binary Issues. *This replicates the previous figure but with the argument corresponding to the log likelihood-ratio pair* $(1, 0)$ *mapped to* $(2, 0)$*.*

Finite State Spaces. A similar kind of intuitive embedding exists for finite state spaces. Fixing a numeraire state, each dimension corresponds to the relative weight between the numeraire state and one of the other states. For example, consider $n = 3$ states under the embedding

$$(p_0, p_1, p_2) \longmapsto \left(w_1 \log \frac{p_1}{p_0},\ w_2 \log \frac{p_2}{p_0} \right).$$

Graphically, the image of $\mathbf{R}^2$ as mapped back into the probability simplex resembles the relationship in the case of two independent binary issues, only with greater warping due to the simplex's triangular shape. All but two directions of arguments (specifically, 180° and 270° counterclockwise from $(1, 0)$) correspond to a unique limit state. As with the case of multiple binary issues, the border directions of the space partition $\mathbf{R}^2$ into regions. However, instead of each direction corresponding to a frequency of discussion between multiple states, each direction captures the relative frequency at which each non-numeraire state is compared with the numeraire.

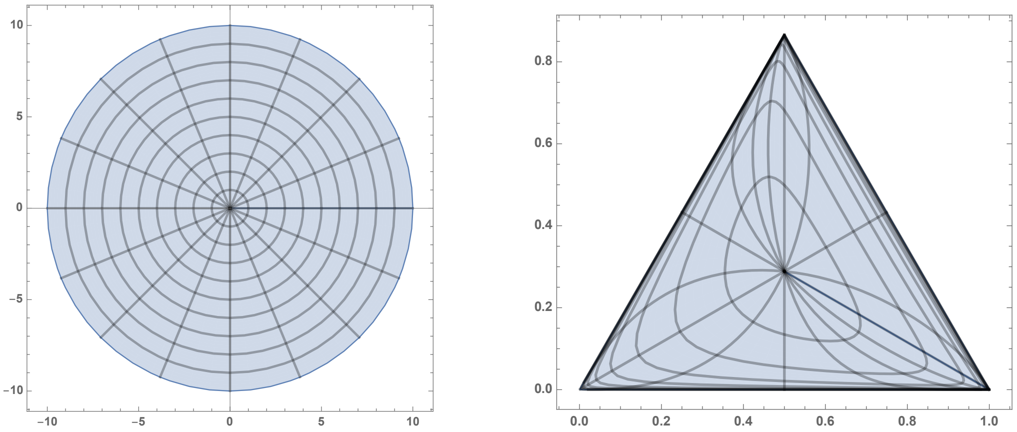

FIGURE 7. Embedding of $\mathcal{A}$ into Δ^2 for Three States. *The left figure shows polar contour lines in $\mathbf{R}^2$; the right figure shows the same contour lines under the following transformation: each argument in $\mathcal{A}$ as embedded in Δ^2 is identified with the posterior distribution it induces on the uniform distribution.*

4.4. **Bayesian Comparison: General Connections**

In general, Euclidean embeddings of Virtual Bayesians need not admit intuitive interpretations with each axis corresponding to an issue or a frequency of discussion. However, the core features of the examples above – that argument length corresponds asymptotically to probabilistic certainty and that argument direction corresponds to a limit state – are preserved in any embedding.

Formally, given a finite state space $\mathcal{X}$, let $\chi \subset \mathcal{X}$ denote any strict and non-empty subset of $\mathcal{X}$: $\chi \notin \{\emptyset, \mathcal{X}\}$. Each possible value of χ corresponds to a region of limit beliefs in which some states of $\mathcal{X}$ have been ruled out. Say that χ_1 and χ_2 are *neighboring* if one is a subset of the other: $\chi_2 \subset \chi_1$ or $\chi_1 \subset \chi_2$. The border of the probability simplex represents all of the limit beliefs a Bayesian agent could adopt. Each vertex, edge, face, etc. of the simplex corresponds to a single χ, and two components of the border are adjacent if and only if they correspond to neighboring subsets of $\mathcal{X}$.

DUALITY BETWEEN ARGUMENT DIRECTIONS AND LIMIT BELIEFS. Now suppose that $(\mathcal{A}, \mathcal{B})$ is a Bayesian learning rule on $\mathcal{X}$. A given subset of states χ is said to be *reached* by an argument $a \in \mathcal{A}$ if repeatedly applying a to any prior belief leads to posteriors which put vanishing weight on any states outside χ and non-vanishing weight on any states in χ. Formally, for any $\pi \in \mathcal{B}$ and any $\varepsilon > 0$, (1) $a^k(\pi)(x) < \varepsilon$ for any $x \notin \chi$ and k sufficiently large and (2) if $x \in \chi$ and $\pi(x) \geq \varepsilon$ then $a^k(\pi)(x) \geq \varepsilon$ for all $k \geq 1$. The Proposition below establishes the connections between values of χ and directions in $\mathbf{R}^n$ for a learning rule that is *full-dimensional*, that is $n = d(\mathcal{A})$. Given a linear extension for $\mathcal{A}$, every value of χ corresponds to a cone of arguments in $\mathbf{R}^n$, or equivalently, a region of vectors on the unit sphere. The association between state subsets and regions of directions in $\mathbf{R}^n$ matches the neighboring relation: neighboring values of χ correspond to adjacent regions of directions.

The critical difference between state subsets and regions of directions lies in their dimensionality. For a state space with $|\mathcal{X}|$ elements, a full-dimensional $(\mathcal{A}, \mathcal{B})$ is extended to $\mathbf{R}^n$ where $n = |\mathcal{X}| - 1$. Then, a region of directions corresponding to any singleton χ has dimension $n-1$; a region corresponding to χ with $|\chi| = 2$ has dimension $n-2$, etc. In this way the geometric measures introduced by the agreement function capture the same *qualitative* information as distance and direction in the probability simplex while stressing different perspectives. For example, if two different arguments $a_1 \neq a_2$ correspond to the same value of a singleton χ, then the sequences $a_1^k(\pi)$ and $a_2^k(\pi)$ both converge to the same point in the probability simplex as $k \longrightarrow \infty$, while the sequences of arguments a_1^k and a_2^k diverge in $\mathbf{R}^n$. Conversely, if χ has $|\mathcal{X}| - 1$ elements, as it rules out only a single state, then there is a only a single direction of arguments which reach χ, while in the probability simplex χ corresponds to a hyperface of the border of $\Delta(\mathcal{X})$.

Proposition 3. *Suppose $(\mathcal{A}, \mathcal{B})$ is a full-dimensional Bayesian learning rule on a finite state space $\mathcal{X}$ and $f : \mathcal{A} \longrightarrow \mathbf{R}^n$ is an essential embedding. Then*

(a) *For each χ there exists a path-connected subset $S(\chi)$ of the unit sphere $S^{|\mathcal{X}|-1}$ such that χ is reached by all arguments a for which $f(a)/|f(a)| \in S(\chi)$.*

(b) *Subsets χ_1 and χ_2 are neighboring if and only if region $S(\chi_1) \cup S(\chi_2)$ is path-connected.*

(c) *The region $S(\chi)$ is a manifold of dimension $|\mathcal{X}| - |\chi| - 1$.*

Proof in the appendix.

5. Methods for Belief Elicitation

Experimenters in the lab are often tasked with eliciting a subject's probabilistic belief about an issue. The machinery of learning rules provides tools for interpreting and quantifying beliefs even when these beliefs are arbitrarily encoded. In essence, under the assumption that a subject's updating is isomorphic to the objective (Bayesian) learning rule, the experimenter can elicit the subject's prior belief by introducing new information and tracking how he updates in response. The techniques discussed below do not exhaust all possible information environments, and they are not the only methods possible in the environments considered; rather, they demonstrate the basic idea of elicitation through *changes* in a subject's beliefs.

An Example. There is a binary question whose answer is known to an experimenter but unknown to a subject. For instance, 'is it currently raining in Seattle?' The subject's belief about the question is 'probably not.' Despite having a grasp of the issue – he recognizes that both states are possible, he knows that rain is generally less common than no rain but that the Pacific Northwest is known for precipitation, he considers the month of the year, etc. – the subject cannot readily distill these factors into a probabilistic assessment. The experimenter takes the following approach. She shows the subject two urns, a 'rain' urn with two-thirds 'rain' balls and one-third 'dry' balls, and a 'dry' urn with the reversed proportions of balls. She selects one urn according to the answer to the question and then draws with replacement from that urn. After each draw, the subject is asked which urn is more likely.

If the subject switches from expressing a belief that no precipitation is most likely to rain being most likely following a sequence of draws with N net rain balls, the experimenter has reached a concrete interpretation of the subject's prior belief. Under the assumption that the subject is a Virtual Bayesian, his prior belief is (approximately) equivalent to starting at the ur-prior of equal weight on both states and receiving $-N$ net rain balls worth of information. In other words, a Bayesian who switched from claiming rain was less likely to more likely at the same time as the subject did would have to had a prior belief of rain with $1/(1 + (1/2)^N)$ probability.

Elicitation Assumptions. Like more traditional approaches, the elicitation method outlined in the example above and detailed below requires stringent assumptions on the agent's behavior. The three assumptions are non-traditional, rendering them useful in some contexts and not others. The first assumption is that the subject's learning rule is isomorphic to a particular Bayesian with identified ur-prior. For instance, in the rain-urn example above, the experimenter assumed the subject's arguments to be isomorphic to a two-state Bayesian with an ur-prior corresponding to the uniform distribution. Second, the subject must be able to report her hypothetical belief after receiving a particular signal. In the example above, the experimenter may have needed to re-shuffle the order of draws shown to the subject in order to induce the subject's belief switch which is necessary for identification, so the subject must agree to treat the draws as though they were strictly i.i.d. More generally, the underlying issue's answer may be unknown to both subject and experimenter, and so the subject has to be able to indicate hypothetical updating.

Finally, the subject must have a rudimentary understanding of the geometric relationships between arguments/beliefs. I consider two possibilities. The first possibility is that the subject can distinguish *relative angles*: he can say, for any triple of arguments/beliefs (a_1, a_2, a_3), whether a_1 is closer in angle to a_2 or to a_3. Relative angle is often a very intuitive concept. In the example above, there are only two directions arguments can take – the 'rain is more likely' direction and the 'rain is less likely' direction – so distinguishing relative angle requires the subject only to state which of the two states is more likely. In the case of Beta learning, each direction corresponds to the mean of the distribution. The second possibility is that the subject distinguishes *relative magnitudes*: he can say which of a_1 and a_2 is a larger argument. This is also an intuitive concept; the experimenter needs only ask the subject, 'putting aside the question of which state these arguments points to, which one incorporates the larger amount of information?' Note that because information is presented in discrete portions, the implied beliefs elicited will necessarily be approximations; presenting the subject with smaller arguments would refine that approximation.

Method via Relative Angles. The example above already illustrates how the experimenter can elicit a subject's belief when his learning rule is isomorphic to a two-state Bayesian. In the case that his learning rule instead matches $\mathbf{Z}_{\geq 0} \times \mathbf{Z}_{\geq 0}$, as with a Beta learner, a simple procedure for identifying his prior requires the following. See Figure 8 for a graphical description. Let the two possible signals that the experimenter can show the subject be denoted x and y, and write a_x and a_y as arguments corresponding to receiving one of those signals. The experimenter posits that the subject begins at prior (b_x, b_y); she asks him whether his belief is closer in angle to a_x or a_y. Without loss of generality, take this to be a_x. Next, the experimenter presents the subject with a_y signals until, following Δ_y signals, he claims his belief is now closer to a_y. Next, she presents the subject with Δ_x of the a_x signals until he claims a_y is equally close to his posterior belief as to his prior belief.

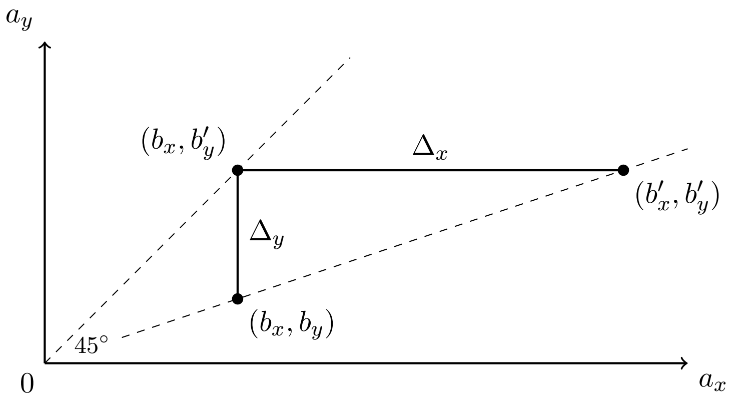

FIGURE 8. Elicitation via Relative Angle. *The experimenter gives the subject Δ_y of the a_y signals to reach the 45° line followed by Δ_x of the a_x signals to return to the angle of the original belief.*

The values of Δ_x and Δ_y identify the subject's prior. First, the fact that Δ_y of the a_y signals equalize the angle between the subject's belief and the two extreme directions implies $b_x = b_y + \Delta_y a_y$. Second, it follows that the subject's ultimate posterior belief is a scaled version of the prior,

$$\frac{b_x + \Delta_x a_x}{b_x} = \frac{b_y + \Delta_y a_y}{b_y}.$$

Together these identities constitute a linear system from which b_x and b_y are recovered.

METHOD VIA RELATIVE MAGNITUDES. When the subject's learning rule is isomorphic to a two-state Bayesian, as in the example, the experimenter can leverage relative magnitude in two steps. She first discerns which of the two signals would make the subject's belief smaller. In the example, for instance, this would be the rain ball. Next, she presents with subject with this signal until, after N iterations, he claims his posterior belief exceeds the magnitude of his prior belief. His prior belief must have been equivalent to $N/2$ of the opposite signal.

When the subject's learning rule is isomorphic to $\mathbf{Z}_{\geq 0} \times \mathbf{Z}_{\geq 0}$, the experimenter first fixes some reference value $\Delta_y > 0$. Next she asks the subject how many x signals would be needed so that his prior plus Δ_y of the a_y arguments would be equivalent in magnitude to Δ_x of the a_x arguments. Finally she asks how many iterations of *both* x and y signals would be necessary to match the magnitude of his belief following Δ_y of the y signals? Denoting this value by Δ, it follows

$$b_x^2 + (b_y + \Delta_y a_y)^2 = (b_x + \Delta_x a_x)^2 + b_y^2 = (b_x + \Delta a_x)^2 + (b_y + \Delta a_y)^2,$$

which similarly produces a linear system identifying b_x and b_y. See Figure 9 for a graphical illustration.

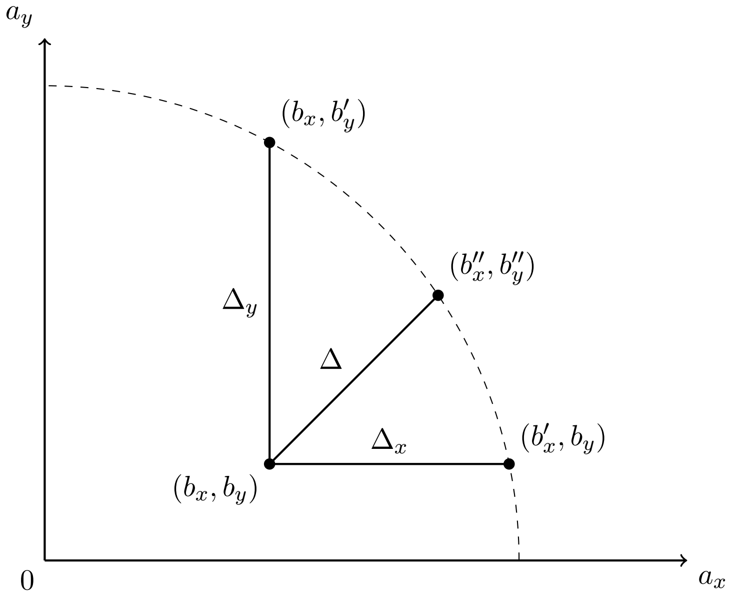

FIGURE 9. Elicitation via Relative Magnitude. *The experimenter gives the subject Δ_y of the a_y signals, Δ_x of the a_x signals, and Δ of both signals such that all three posteriors have the same magnitude.*

6. Discussion

This paper has analyzed Bayesian updating from algebraic and geometric perspectives, starting from the key observation that Bayesians treat signal realizations, here codified as *arguments*, like elements of real vector spaces. Hence, updating rules which share the compositional properties of Bayesians are characterized by the substantive axioms of Theorem 1 – injectivity, commutativity, acyclicality – which allow an embedding into $\mathbf{R}^n$. As Theorem 3 establishes, there is a one-to-one relationship between embeddings and additive agreement functions. Moreover, any such updating rule inherits the geometric structures of Euclidean space via its embedding, with measures of length and angle that provide a prior-free understanding of the relationships between pieces of information. As Section 5 illustrates, these conceptual tools help to engage agents who lack probabilistic sophistication but nonetheless maintain Bayesian-like internal consistency.

In what follows, I discuss what alterations to the modeling framework would have led to a substantially different characterization theorem and outline several open questions which could be addressed in subsequent work.

6.1. Discussion of Modeling Assumptions

To understand how different modeling assumptions could have led to substantively different characterization theorems, note first that the framework of learning rules largely serves as a notational container for richer structures. Rather, it is the regularity conditions placed on Bayesian learning rules which are most restrive. The non-dogmatic learning requirement ensures Bayes' rule is always applicable, as the belief set contains no elements which rule out the arrival of any arguments, but it also excludes a number of examples which would be regularly termed 'Bayesian.' The challenge is this: if a Bayesian is allowed to receive an argument which tells him some state $x^* \in \mathcal{X}$ should have zero probability, how should his belief transition after subsequently receiving another argument that tells him to put *more* weight on state x^*? There are multiple paths to follow here, but none is clearly best. To borrow from game theory, the logic of Perfect Bayesian Equilibrium suggests the agent could have any belief at all, while that of Sequential Equilibrium would suggest introducing a topology on the beliefs and arguments to define the agent's update as the limit of Bayesian prescriptions. A third route might involve the agent 'forgetting' the earlier information which caused him to rule out x^*, but this would break commutativity and could therefore radically change the nature of Bayesianism.

Although the *i.i.d.* assumption might seem strong, it is not substantive. One could alternatively omit X from the model and specify a Bayesian learner solely with reference to the probability space and the sequence $(Y_i)_{i=1}^{\infty}$, which is then assumed to be *exchangeable.* Instead of having belief states in the form of conditional assessments about the distribution of X, the leaner's beliefs concern the distribution of the tail of the sequence, $(Y_i)_{i=t+1}^{\infty}$. However, by the de Finetti–Hewitt–Savage theorem, the distribution over $(Y_i)_{i=1}^{\infty}$ can equivalently be expressed as a joint distribution over $(Y_i)_{i=1}^{\infty}$ and an underlying measure, $F \in \Delta(\mathcal{Y})$, conditional on which the $(Y_i)_{i=1}^{\infty}$ sequence is i.i.d. Then, assessments about the tail distribution are equivalent to assessments about the measure F, which plays the same role in the model as the X does in the version above. So while the omission of X and 'generalization' to exchangeable Y_i's may seem at first more elegant, this route only leads back, with additional trouble, to the model as presented in the main text.

The assumption that the signal space is discrete and the state space is either discrete or continuous sidelines a host of measure-theoretic complications. It is unclear how exactly Theorem 1 might change as a result of relaxing that, but I do not suspect any dramatic difference. There are a host of examples with a continuous signal space, e.g. the Gaussian-Gaussian learning model, that nonetheless satisfy all but the countability axiom and embed neatly into $\mathbf{R}^n$.

6.2. Future Work

Looking forward, this paper could provide a foundation for additional applications of the learning rule framework. For example, Theorems 1 and 2 show how to link an arbitrary learning rule with *some* Bayesian, but they do not address the question of how many Bayesian isomorphism classes exist, which could be addressed in subsequent work.[13] Along similar lines, non-Bayesians could be further classified in terms of their sophistication vis-a-vis Bayesians. Say that learning rule $(\mathcal{A}', \mathcal{B}')$ *computes for*, or reproduces, $(\mathcal{A}, \mathcal{B})$ if there exists an injective homomorphism $(\mathcal{A}, \mathcal{B}) \longrightarrow (\mathcal{A}', \mathcal{B}')$. This defines a complete and transitive relation on the set of all learning rules and yield a partially ordered set of computability equivalence classes. What characterizes the non-Bayesians which can compute for some or for all Bayesians, and what non-Bayesians can Bayesians compute for?

More broadly, the geometric concepts of argument scaling, projecting one argument onto another, or squeezing two arguments nearer by reducing the angle between them could be used as tools for modeling cognitive biases. Indeed, in a recent working paper (Chauvin, 2020) I have developed a model of an agent who subjectively distorts information by scaling multiple arguments in the interest of maximizing the magnitude of their composition. Furthermore, the framework supplies a novel model of naive updating in networks: suppose a network of agents start from a common prior, are initially supplied with a single private argument each, then update by treating their neighbors' opinions as reflecting new arguments. Under intuitive circumstances,[14] the beliefs from this *aggregation* process coincide with the *averaging* of DeGroot learning. In other cases, such as reasoning over a binary state, repeated aggregation leads to consensus on one of the two states, contrasting with DeGroot's prediction of an interior probabilistic consensus. Future work could compare the two models in greater detail.

[13] It is relatively easy to see that there is a family of Bayesian examples for $\mathbf{Z}^{n_1} \times \mathbf{Z}^{n_1}_{\geq 0} \times \mathbf{Z}^{n_3}_{>0}$ and $\mathbf{Q}^{n_1} \times \mathbf{Q}^{n_1}_{\geq 0} \times \mathbf{Q}^{n_3}_{>0}$ for almost all specifications of n_1, n_2, n_3, and that these are all non-isomorphic. How many others are there?

[14] Namely, when the data generating process is Gaussian, when all agents have the same number of neighbors, and when agents place equal weight on their neighbors.

References

Abreu, Dilip and Ariel Rubinstein, "The Structure of Nash Equilibrium in Repeated Games with Finite Automata," *Econometrica*, 1988, *56* (6), 1259–1281.

Benjamin, Daniel J., Aaron Bodoh-Creed, and Matthew Rabin, "Base-Rate Neglect: Foundations and Implications," *Working Paper*, 2019.

Camerer, Colin F. and Teck-Hua Ho, "Experience-Weighted Attraction Learning in Normal Form Games," *Econometrica*, 1999, *67* (4), 827–874.

Chauvin, Kyle P., "Belief Updating with Dissonance Reduction," *Working Paper*, 2020.

Cripps, Martin W., "Divisible Updating," *Working Paper*, 2019.

DeGroot, Morris H., "Reaching a Consensus," *Journal of the American Statistical Association*, 1974, *69* (345), 118–121.

Epstein, Larry G., "An Axiomatic Model of Non-Bayesian Updating," *The Review of Economic Studies*, 2006, *73*, 413–436.

________**, Jawwad Noor, and Alvaro Sandroni**, "Non-Bayesian Updating: a Theoretical Framework," *Theoretical Economics*, 2008, *3*, 193–229.

________ **and Kyoungwon Seo**, "Symmetry of Evidence Without Evidence of Symmetry," *Theoretical Economics*, 2010, *5* (3), 313–368.

Erev, Ido and Alvin E. Roth, "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria," *American Economic Review*, 1998, *88* (4), 848–881.

Esponda, Ignacio and Demian Pouzo, "Berk–Nash Equilibrium: A Framework for Modeling Agents With Misspecified Models," *Econometrica*, 2016, *84* (3), 1093–1130.

Eyster, Erik and Matthew Rabin, "Naive Herding in Rich-Information Settings," *American Economic Journal: Microeconomics*, 2010, *2* (4), 221–43.

Frankel, Alexander and Emir Kamenica, "Quantifying Information and Uncertainty," *American Economic Review*, 2019, *109* (10), 3650–3680.

Fuchs, László, *Abelian Groups*: Springer International Publishing, 2015.

Hanany, Eran and Peter Klibanoff, "Updating Ambiguity Averse Preferences," *The BE Journal of Theoretical Economics*, 2014, *9* (1), 1935–1704.

Lehrer, Ehud and Roee Teper, "Who is a Bayesian," *Working Paper*, 2016.

Molavi, Pooya, Alireza Tahbaz-Salehi, and Ali Jadbabaie, "A Theory of Non-Bayesian

Social Learning," *Econometrica*, 2018, *86* (2), 445–490.

Prelec, Drazen, "The Probability Weighting Function," *Econometrica*, 1998, *66* (3), 497.

Rabin, Matthew and Joel L. Schrag, "First Impressions Matter: A Model of Confirmatory Bias," *Quarterly Journal of Economics*, 1999, *114* (1), 37–82.

Reddy, Uday S., "Notes on Semigroups," *Unpublished Manuscript*, 2014.

Schlag, Karl H., James Tremewan, and Joël J. van der Weele, "A Penny for Your Thoughts: a Survey of Methods for Eliciting Beliefs," *Experimental Economics*, 2015, *18*, 457–490.

Schotter, Andrew and Isabel Trevino, "Belief Elicitation in the Laboratory," *Annual Review of Economics*, 2014, *6* (1), 103–128.

Shmaya, Eran and Leeat Yariv, "Experiments on Decisions under Uncertainty: A Theoretical Framework," *American Economic Review*, 2016, *106* (7), 1775–1801.

Wilson, Andrea, "Bounded Memory and Biases in Information Processing," *Econometrica*, 2014, *82* (6), 2257–2294.

Zhao, Chen, "Pseudo-Bayesian Updating," *Working Paper*, 2016.

Appendix A. Proofs

Proof of Fact 1. The *only if* direction of the proof is immediate from the definition of a learning rule isomorphism. To establish the *if* direction, first define the bijection $f : \mathcal{A} \longrightarrow \mathcal{B}$ by setting $f(a)$ as the unique belief with $a = a_{f(a)}$, and note that $a_2(f(a_1)) = f(a_1 \circ a_2)$ for all $a_1, a_2 \in \mathcal{A}$. Define $f' : \mathcal{A}' \longrightarrow \mathcal{B}'$ accordingly. As

$$f(f^{-1}(b) \circ a) = a(f(f^{-1}((b))) = a(b),$$

applying f^{-1} to both sides yields the relation $f^{-1}(b) \circ a = f^{-1}(a(b))$. Now let $g : \mathcal{A} \longrightarrow \mathcal{A}'$ be a semigroup isomorphism. Define $h : \mathcal{B} \longrightarrow \mathcal{B}'$ by $h(b) = f'(g(f^{-1}(b)))$, and note that h, as a composition of bijections, is itself a bijection. Then for any $a \in \mathcal{A}$ and $b \in \mathcal{B}$,

$$\begin{aligned} g(a)(h(b)) = g(a)(f'(g(f^{-1}(b)))) = f'(g(f^{-1}(b)) \circ g(a)) &= f'(g(f^{-1}(b) \circ a)) \\ &= f'(g(f^{-1}(a(b)))) = h(a(b)), \end{aligned}$$

where the fourth equality holds by the relation established above. This demonstrates that (g, h) is a learning rule isomorphism. ○

Proof of Lemma 1. We show the contrapositive: a case of $a_1(b) = a_2(b)$ where $a_1 \neq a_2$ necessarily implies a violation of (A.2). Towards a contradiction, suppose $a_1(b) = a_2(b)$ for some $a_1 \neq a_2 \in \mathcal{A}$ and $b \in \mathcal{B}$, and let $f : \mathcal{A} \longrightarrow \mathcal{B}$ denote the map defined in the proof of Fact 1 above. Now define $a = f^{-1}(b)$ and $b_1 = f(a_1) \neq b_2 = f(a_2)$. Then by the properties of f and commutativity we have

$$f(a \circ a_1) = a_1(f(a)) = a_1(b) = a_2(b) = a_2(f(a)) = f(a \circ a_2)$$

and

$$\begin{aligned} a(b_1) = a(f(a_1)) = f(a_1 \circ a) &= f(a \circ a_1) \\ &= f(a \circ a_2) = f(a_2 \circ a) = a(g(a_2)) = a(b_2), \end{aligned}$$

contradicting (A.2). ○

Proof of Lemma 2. Let[15] $S \equiv \bar{\mathcal{A}} \times \bar{\mathcal{A}}$, and define the addition of $s^1 = (a_1^1, a_2^1)$ and $s^2 = (a_1^2, a_2^2)$ by $s^1 + s^2 = (a_1^1 \circ a_1^2, a_1^2 \circ a_2^2)$. Then define the equivalence relation $(a_1^1, a_2^1) \sim (a_1^2, a_2^2)$ if $a_1^1 \circ a_2^2 = a_1^2 \circ a_2^1$, and let $\mathcal{A}^+ = S/\sim$ be the set of equivalence classes of S under $\sim$, with typical members denoted $\bar{s}$. Define the composition of $\bar{s}^1$ and $\bar{s}^2$ to be the equivalence class containing $s^1 + s^2$, where $s^1 \in \bar{s}^1$ and $s^2 \in \bar{s}^2$ are any representative members of S. To verify this operation is well defined, suppose that $s^1, \hat{s}^1 \in \bar{s}^1$ and $s^2, \hat{s}^2 \in \bar{s}^2$. Then

$$\begin{cases} s^1 \sim \hat{s}^1 \implies a_1^1 \circ \hat{a}_2^1 = \hat{a}_1^1 \circ a_2^1 \\ s^2 \sim \hat{s}^2 \implies a_1^2 \circ \hat{a}_2^2 = \hat{a}_1^2 \circ a_2^2. \end{cases}$$

Combining these equations and leveraging commutativity,

$$a_1^1 \circ a_1^2 \circ \hat{a}_2^1 \circ \hat{a}_2^2 = \hat{a}_1^1 \circ \hat{a}_1^2 \circ a_2^1 \circ a_2^2,$$

confirming

$$s^1 + s^2 = (a_1^1 \circ a_1^2, a_2^1 \circ a_2^2) \sim (\hat{a}_1^1 \circ \hat{a}_1^2, \hat{a}_2^1 \circ \hat{a}_2^2) = \hat{s}^1 + \hat{s}^2.$$

We now establish that $\mathcal{A}^+$ is an Abelian group. Closure under addition, associativity, and commutativity follow from $\bar{\mathcal{A}}$. The equivalence class containing all (a, a) is an identity element. The class corresponding to $-s \equiv (a_2, a_1)$ is the inverse of that containing $s = (a_1, a_2)$. Furthermore, for every $a \in \bar{\mathcal{A}}$, there are at most $|\bar{\mathcal{A}}|$ corresponding equivalence classes in $\mathcal{A}^+$. Thus from $\bar{\mathcal{A}}$ countable we conclude $\mathcal{A}^+$ is countable.

Next we verify the map $\bar{f} : \bar{\mathcal{A}} \longrightarrow \mathcal{A}^+$, $a \longmapsto (a, a_0)$ is a homomorphism: $\bar{f}(a_1 + a_2) = (a_1 + a_2, a_0) = (a_1, a_0) + (a_2, a_0) = \bar{f}(a_1) + \bar{f}(a_2)$. It is also injective, as

$$\bar{f}(a_1) = \bar{f}(a_2) \implies (a_1, a_0) = (a_2, a_0) \implies a_1 = a_1 \circ a_0 = a_2 \circ a_0 = a_2.$$

Thus the composition $f : \mathcal{A} \longrightarrow \bar{\mathcal{A}} \longrightarrow \mathcal{A}^+$ is an injective homomorphism. Finally, note that an equivalence class $\bar{s} \in \mathcal{A}^+$ containing (a_1, a_2) is effectively the composition of a_1 with the 'inverse' of a_2, as $\bar{s}$ also contains $(a_1, a_0) \circ (a_0, a_2)$. Therefore, in the notation of the inclusion map, $\bar{s} = f(a_1) \circ (-f(a_2))$. $\circ$

[15] N.b. the main content of this Lemma is an established result from abstract algebra. I nonetheless include a full proof so that interested readers can follow the complete construction.

Proof of Lemma 3. We make use of two key facts from group theory:

(1) If $\mathcal{A}^+$ is an Abelian group, then there exists a divisible group D (an Abelian group such $D = nD$ for all integers $n \geq 1$) and an embedding $h : \mathcal{A}^+ \longrightarrow D$ as an essential subgroup: for all non-trivial subgroups $H \subset D$, $H \neq \{0\}$, the image of $\mathcal{A}^+$ in D intersects H: $h(\mathcal{A}^+) \cap H \neq \{0\}$.

(2) Any acyclic (equivalently, in the parlance of group theory, 'torsion-free') divisible group D is isomorphic to a direct sum of copies of $\mathbf{Q}$.[16]

Note that as $\mathcal{A}^+$ is acyclic, its divisible counterpart D must be as well: were there any $x \in D$, $x^n = 0$, then $H \equiv \{d^n | n \geq 0\}$ is a non-trivial subgroup, which by the essentialness of the embedding implies $a^* = h^{-1}(x^d) \in \mathcal{A}^+$ is an element with $(a^*)^m = a_0$ for some m, contradicting the acyclicality of $\mathcal{A}^+$. By these facts, it follows there exists an embedding $f : \mathcal{A}^+ \longrightarrow \mathcal{A}^* \cong \mathbf{Q}^{d(\mathcal{A})}$ for some value (cardinality) $d(\mathcal{A})$. Let $(\bar{a}_i)_{i=1}^{d(\mathcal{A})}$ be a basis for $\mathcal{A}^*$. As $f(\mathcal{A}^+)$ is an essential subgroup of $\mathcal{A}^*$, it shares a non-trivial intersection with the subgroup generated by each basis element $\bar{a}_i$, and thus there exists $q_i \in \mathbf{Q}$ such that $q_i \bar{a}_i \in f(\mathcal{A}^+)$. Define $\hat{a}_i = q_i \bar{a}_i$. Then $(\hat{a}_i)_{i=1}^{d(\mathcal{A})}$ is a basis for $\mathcal{A}^*$: each element $a \in \mathcal{A}^*$ can be uniquely written $a = \sum_{i=1}^{d(\mathcal{A})} c_i(a) \bar{a}_i$, so

$$a = \sum_{i=1}^{d(\mathcal{A})} \underbrace{\frac{1}{q_i} c_i(x)}_{\hat{c}_i(x)} \hat{a}_i$$

is an equivalent unique representation as a rational combination of $\hat{a}_i$'s. This reasoning also demonstrates that $d(\mathcal{A})$ can be no greater than the cardinality of $\mathcal{A}^+$, which is countable. ○

Proof of Lemma 4. Let $\mathcal{X} = \{0, 1\}$, let $\mathcal{Y} = \mathcal{A}$, and denote the homomorphism $g : \mathcal{A} \longrightarrow \mathbf{R}$. Since g is two-sided, there exists $a_1, a_2 \in \mathcal{A}$ with $g(a_1) > 0 > g(a_2)$. In what follows we find a joint distribution over (X, Y_i), $X \in \mathcal{X}$, $Y_i \in \mathcal{Y}$, such that the log likelihood-ratio function associated with each $a \in \mathcal{A}$ is equal to $g(a)$. This produces a Bayesian learning rule which is isomorphic to its associated additive semigroup of log-likelihood functions, and therefore also isomorphic to $\mathcal{A}$, verifying $\mathcal{A}$ as a Virtual Bayesian. To accomplish this, let the marginal probabilities of X be set to $\mathbf{P}[X = 1] = \mathbf{P}[X = 0] = 1/2$, and note that, given a marginal distribution over $\mathcal{Y} = \mathcal{A}$, the

[16] See, for example, Fuchs (2015) pp. 131-141.

conditional probabilities must satisfy

$$\log\left(\frac{\mathbf{P}[y=a|X=1]}{\mathbf{P}[y=a|X=0]}\right)=g(a)\implies\begin{cases}\mathbf{P}[y=a|X=1]=\dfrac{\exp\{g(a)\}}{2(1+\exp\{g(a)\})}\mathbf{P}[y=a]\\ \mathbf{P}[y=a|X=0]=\dfrac{1}{2(1+\exp\{g(a)\})}\mathbf{P}[y=a].\end{cases}$$

It thus suffices to find a marginal distribution over $\mathcal{Y}$ that satisfies the law of total probability

$$\mathbf{P}[X=1]=\frac{1}{2}=\sum_{a\in\mathcal{A}}\frac{1}{1+\exp\{-g(a)\}}\mathbf{P}[y=a]=\sum_{a\in\mathcal{A}}\mathbf{P}[X=1|y=a]\mathbf{P}[y=a].$$

Define $E:\Delta^{\circ}(\mathcal{A})\longrightarrow(0,1)$ by

$$\delta\longmapsto\sum_{a\in\mathcal{A}}\frac{\delta(a)}{1+\exp\{-g(a)\}},$$

and note that $\Delta^{\circ}(\mathcal{A})$ is a connected metric space, that E is continuous and that, as $g(a_1)>0>g(a_2)$, it follows $E(\delta_{a_1})>1/2>E(\delta_{a_2})$ for distributions δ_{a_1} and δ_{a_2} that place sufficiently high mass on a_1 and a_2 respectively. By the intermediate value theorem, there exists δ^* such that $E(\delta^*)=1/2$. This completes the proof. ○

Proof of Lemma 5. Let $(\mathcal{A},\mathcal{B})$ be the Bayesian learning rule corresponding to $\{X,(Y_i)_{i=1}^{\infty}\}$ on $(\Omega,\mathcal{F},\mathbf{P})$. Countability of $\mathcal{A}$ and $\mathcal{B}$ follow from the assumed countability of $\mathcal{Y}$. $(\mathcal{A},\mathcal{B})$ is self-recording as the belief $\mathbf{P}_X[\cdot]$ is an ur-prior.

To show all $a\in\mathcal{A}$ are injective, let $\mathbf{P}_X[\cdot|y_1,\ldots,y_n]$ and $\mathbf{P}_X[\cdot|y_1',\ldots,y_n']$ be two distinct beliefs. Then there exists some values $x^+,x^-\in\mathcal{X}$ such that

$$\frac{\mathbf{P}[x^+|y_1,\ldots,y_n]}{\mathbf{P}[x^-|y_1,\ldots,y_n]}>\frac{\mathbf{P}[x^+|y_1',\ldots,y_n']}{\mathbf{P}[x^-|y_1',\ldots,y_n']}.$$

Then for any $\hat{y}_1,\ldots,\hat{y}_m$,

$$\begin{aligned}\frac{\mathbf{P}[x^+|y_1,\ldots,y_n,\hat{y}_1,\ldots,\hat{y}_m]}{\mathbf{P}[x^-|y_1,\ldots,y_n,\hat{y}_1,\ldots,\hat{y}_m]}&=\frac{\mathbf{P}[x^+|y_1,\ldots,y_n]}{\mathbf{P}[x^-|y_1,\ldots,y_n]}\cdot\frac{\mathbf{P}[\hat{y}_1,\ldots,\hat{y}_m|x^+]}{\mathbf{P}[\hat{y}_1,\ldots,\hat{y}_m|x^-]}\\&>\frac{\mathbf{P}[x^+|y_1',\ldots,y_n']}{\mathbf{P}[x^-|y_1',\ldots,y_n']}\cdot\frac{\mathbf{P}[\hat{y}_1,\ldots,\hat{y}_m|x^+]}{\mathbf{P}[\hat{y}_1,\ldots,\hat{y}_m|x^-]}=\frac{\mathbf{P}[x^+|y_1',\ldots,y_n',\hat{y}_1,\ldots,\hat{y}_m]}{\mathbf{P}[x^-|y_1',\ldots,y_n',\hat{y}_1,\ldots,\hat{y}_m]},\end{aligned}$$

so $\mathbf{P}_X[\cdot|y_1,\dots,y_n,\hat{y}_1,\dots,\hat{y}_m]$ and $\mathbf{P}_X[\cdot|y_1',\dots,y_n',\hat{y}_1,\dots,\hat{y}_m]$ are also distinct beliefs. Commutativity comes directly from the assumption of conditional independence:

$$\begin{aligned}\mathbf{P}[x|y_1,\dots,y_n,\hat{y}_1,\dots,\hat{y}_m] &= \frac{\mathbf{P}[y_1,\dots,y_n,\hat{y}_1,\dots,\hat{y}_m|x]\mathbf{P}[x]}{\sum_{x'\in\mathcal{X}}\mathbf{P}[y_1,\dots,y_n,\hat{y}_1,\dots,\hat{y}_m|x']\mathbf{P}[x']}\\ &= \frac{\prod_{i=1}^n \mathbf{P}[y_i|x]\prod_{j=1}^m \mathbf{P}[\hat{y}_j|x]\mathbf{P}[x]}{\sum_{x'\in\mathcal{X}}\prod_{i=1}^n \mathbf{P}[y_i|x']\prod_{j=1}^m \mathbf{P}[\hat{y}_j|x']\mathbf{P}[x']}\\ &= \frac{\mathbf{P}[\hat{y}_1,\dots,\hat{y}_m,y_1,\dots,y_n|x]\mathbf{P}[x]}{\sum_{x'\in\mathcal{X}}\mathbf{P}[\hat{y}_1,\dots,\hat{y}_m,y_1,\dots,y_n|x']\mathbf{P}[x']} = \mathbf{P}[x|\hat{y}_1,\dots,\hat{y}_m,y_1,\dots,y_n].\end{aligned}$$

To establish $(\mathcal{A},\mathcal{B})$ as acyclic, let $a = a_{(\hat{y}_1,\dots,\hat{y}_m)}$ be any non-identity argument. As it is not the identity, there exists some belief $\mathbf{P}_X[\cdot|y_1,\dots,y_n]$ and some values $x^+,x^-\in\mathcal{X}$ such that

$$\frac{\mathbf{P}[x^+|y_1,\dots,y_n,\hat{y}_1,\dots,\hat{y}_m]}{\mathbf{P}[x^-|y_1,\dots,y_n,\hat{y}_1,\dots,\hat{y}_m]} = \frac{\mathbf{P}[x^+|y_1,\dots,y_n]}{\mathbf{P}[x^-|y_1,\dots,y_n]}\cdot\underbrace{\frac{\mathbf{P}[\hat{y}_1,\dots,\hat{y}_m|x^+]}{\mathbf{P}[\hat{y}_1,\dots,\hat{y}_m|x^-]}}_{>0} > \frac{\mathbf{P}[x^+|y_1,\dots,y_n]}{\mathbf{P}[x^-|y_1,\dots,y_n]},$$

and thus the odds-ratio of x^+ to x^- after receiving $k\geq 1$ copies of argument $a_{(\hat{y}_1,\dots,\hat{y}_m)}$ is

$$\frac{\mathbf{P}[x^+|y_1,\dots,y_n]}{\mathbf{P}[x^-|y_1,\dots,y_n]}\cdot\left(\frac{\mathbf{P}[\hat{y}_1,\dots,\hat{y}_m|x^+]}{\mathbf{P}[\hat{y}_1,\dots,\hat{y}_m|x^-]}\right)^k > \frac{\mathbf{P}[x^+|y_1,\dots,y_n]}{\mathbf{P}[x^-|y_1,\dots,y_n]},$$

demonstrating that $a_{(\hat{y}_1,\dots,\hat{y}_m)}$ cannot be cyclic.

Finally we show that $(\mathcal{A},\mathcal{B})$ is pluralistic. Note that for any subset of states $E\subset\mathcal{X}$,

$$\mathbf{P}[E] = \sum_{y\in\mathcal{Y}}\mathbf{P}[E|y]\cdot\mathbf{P}[E],$$

and, for all $y\in\mathcal{Y}$,

$$\mathbf{P}[E|y]+\mathbf{P}[E^c|y] = 1.$$

By non-triviality it must be the case that $\mathbf{P}[E]\neq\mathbf{P}[E|y]$ for some E and y; by the first identity there then exists E,y_1,y_2 such that $\mathbf{P}[E|y_1]>\mathbf{P}[E]>\mathbf{P}[E|y_2]$. It then follows from the second identity that there must then exist E_1,E_2,y_1,y_2 such that

$$\begin{cases}\mathbf{P}[E_1|y_1]>\mathbf{P}[E_1]>\mathbf{P}[E_1|y_2]\\ \mathbf{P}[E_2|y_1]<\mathbf{P}[E_2]<\mathbf{P}[E_2|y_2]\end{cases}\implies \frac{\mathbf{P}[E_1|y_1]}{\mathbf{P}[E_2|y_1]}>\frac{\mathbf{P}[E_1]}{\mathbf{P}[E_2]}>\frac{\mathbf{P}[E_1|y_2]}{\mathbf{P}[E_2|y_2]}.$$

Bayes' rule and the conditional independence assumption then allow us to manipulate this as follows:

$$\frac{\mathbf{P}[y_1|E_1]\mathbf{P}[E_1]}{\mathbf{P}[y_1|E_2]\mathbf{P}[E_2]} > \frac{\mathbf{P}[E_1]}{\mathbf{P}[E_2]} > \frac{\mathbf{P}[y_2|E_1]\mathbf{P}[E_1]}{\mathbf{P}[y_2|E_2]\mathbf{P}[E_2]}$$
$$\frac{\mathbf{P}[y_1|E_1]}{\mathbf{P}[y_1|E_2]} > 1 > \frac{\mathbf{P}[y_2|E_1]}{\mathbf{P}[y_2|E_2]}$$
$$\left(\frac{\mathbf{P}[y_1|E_1]}{\mathbf{P}[y_1|E_2]}\right)^k > 1 > \left(\frac{\mathbf{P}[y_2|E_1]}{\mathbf{P}[y_2|E_2]}\right)^j$$
$$\frac{\mathbf{P}[y_1|E_1]^k\mathbf{P}[E_1]}{\mathbf{P}[y_1|E_2]^k\mathbf{P}[E_2]} > \frac{\mathbf{P}[E_1]}{\mathbf{P}[E_2]} > \frac{\mathbf{P}[y_2|E_1]^j\mathbf{P}[E_1]}{\mathbf{P}[y_2|E_2]^j\mathbf{P}[E_2]}$$
$$\frac{\mathbf{P}[E_1|\overbrace{y_1,y_1,\ldots,y_1}^{k\text{ times}}]}{\mathbf{P}[E_2|\underbrace{y_1,y_1,\ldots,y_1}_{k\text{ times}}]} > \frac{\mathbf{P}[E_1]}{\mathbf{P}[E_2]} > \frac{\mathbf{P}[E_1|\overbrace{y_2,y_2,\ldots,y_2}^{j\text{ times}}]}{\mathbf{P}[E_2|\underbrace{y_2,y_2,\ldots,y_2}_{j\text{ times}}]},$$

where k and j are positive integers, and the last line follows from conditional independence. This demonstrates that $a_{(y_1)}$ and $a_{(y_2)}$ cannot share any common multiples. Hence $\mathcal{A}$ is pluralistic. ○

Proof of Proposition 1. That deductive learning rules are commutative and idempotent is clear by inspection. Now, suppose learning rule $(\mathcal{A}, \mathcal{B})$ is self-recording, countable, commutative, and idempotent. To show that it is isomorphic to some deductive learning rule it suffices to exhibit a set $\mathcal{X}$ such that each $a \in \mathcal{A}$ is associated with $X_a \subset \mathcal{X}$ and $a_1 \circ a_2 = a_3$ if and only if $X_{a_1} \cup X_{a_2} = X_{a_3}$. To that end, define

$$\mathcal{X} \equiv \left\{A \subset \mathcal{A} \mid a \in A,\ a \circ a' \circ a'' = a' \circ a'' \implies a' \in A \text{ or } a'' \in A\right\}$$

and denote $X_a = \{A \in \mathcal{X} \mid a \in A\}$.

We first note two features of the above definition. First, setting $a' = a''$ in the condition for $A \in \mathcal{X}$ produces the simpler condition $a \in A$, $a \circ a' = a' \implies a' \in A$. Second, if $a \neq a'$, then $X_a \neq X_{a'}$. To see this, posit $a \neq a'$, which implies $a \circ a' \neq a$ or $a \circ a' \neq a'$; without loss of generality, take the former as true. Then define $A^* = \{\hat{a} \in \mathcal{A} \mid a \circ \hat{a} \neq a\}$ to be the set of all arguments not 'contained' in a. Next we show $A^* \in \mathcal{X}$: if this were not true, then there would exist $\hat{a} \in A^*$, $\hat{a}', \hat{a}'' \in \mathcal{A}$ for which $\hat{a} \circ \hat{a}' \circ \hat{a}'' = \hat{a}' \circ \hat{a}''$ but $\hat{a}' \notin A^*$ and $\hat{a}'' \notin A^*$; by definition it follows $a \circ \hat{a}' = a$

and $a \circ \hat{a}'' = a$, so $a \circ (\hat{a}' \circ \hat{a}'') = a$, so

$$a \circ \hat{a} = (a \circ \hat{a}' \circ \hat{a}'') \circ \hat{a} = a \circ (\hat{a}' \circ \hat{a}'' \circ \hat{a}) = a \circ \hat{a}' \circ \hat{a}'' = a,$$

contradicting $\hat{a} \in A^*$. As $a' \in A^*$ and $A^* \in \mathcal{X}$, it follows $A^* \in X_{a'}$. On the other hand, $a \notin A^*$, so $A^* \notin X_a$, completing the proof that $X_a \neq X_{a'}$.

We now verify $a_1 \circ a_2 = a_3$ if and only if $X_{a_1} \cup X_{a_2} = X_{a_3}$. Suppose $a_1 \circ a_2 = a_3$. If $A \in X_{a_1}$, then by definition $a_1 \in A$; as $a_1 \circ a_3 = a_3$, it follows $a_3 \in A$, so $A \in X_{a_3}$. Hence $X_{a_1} \subset X_{a_3}$, and (as $X_{a_2} \subset X_{a_3}$ by identical logic) $X_{a_1} \cup X_{a_2} \subset X_{a_3}$. If $A \in X_{a_3}$, then $a_3 \in A$; as $a_3 \circ a_1 \circ a_2 = a_1 \circ a_2$, then either $a_1 \in A \implies A \in X_{a_1}$ or $a_2 \in A \implies A \in X_{a_2}$. Hence $X_{a_3} \subset X_{a_1} \cup X_{a_2}$, so $X_{a_1} \cup X_{a_2} = X_{a_3}$. Finally, as we have shown that $X_{a_1} \cup X_{a_2} = X_{a_1 \circ a_2}$ and that $a_1 \neq a_2 \implies X_{a_1} \neq X_{a_2}$, it follows that $X_{a_1} \cup X_{a_2} = X_{a_3}$ must imply $a_1 \circ a_2 = a_3$.

$\circ$

Proof of Proposition 2. Suppose $(\mathcal{A}, \mathcal{B})$ is a finite, self-recording learning rule satisfying A.3, A.4, A.5, and A.6. By the proof of Theorem 1, $\mathcal{A}$ satisfies the cancellation property. For any $a \in \mathcal{A}$, by finiteness there exist $k, j \geq 1$ such that $a^k = a^{k+j}$, so

$$a^k \circ a^j = a^k \circ a_0,$$

which by the cancellation property implies $a^j = a_0$, so

$$a^{j-1} \circ a = a \circ a^{j-1} = a_0,$$

so $a^{j-1} = a^{-1}$. This makes $\mathcal{A}$ a finite, Abelian group. By the Fundamental Theorem of Finitely Generated Abelian Groups, $\mathcal{A} \cong \mathbf{Z}_{p_1} \times \cdots \times \mathbf{Z}_{p_k}$. Moreover, as $(\mathcal{A}, \mathcal{B})$ is pluralistic, $k \geq 2$. If $(\mathcal{A}, \mathcal{B})$ is a finite, self-recording learning rule with $\mathcal{A} \cong \mathbf{Z}_{p_1} \times \cdots \times \mathbf{Z}_{p_k}$, $k \geq 2$, it is readily verified that it satisfies axioms (A.3,4,5,6).

$\circ$

Proof of Theorem 2. We begin by establishing the following lemma.

Lemma. *Let γ be an additive agreement function on $\mathcal{A}$. Then*

(a) *γ extends uniquely to an additive agreement function $\gamma^* : \mathcal{A}^* \times \mathcal{A}^* \longrightarrow \mathbf{R}$.*

(b) *γ identifies $\mathcal{A}$: if $\gamma(a_1, a) = \gamma(a_2, a)$ for all $a \in \mathcal{A}$, then $a_1 = a_2$.*

To prove statement (a), we first show that γ admits an extension to $\mathcal{A}^+ \times \mathcal{A}^+ \longrightarrow \mathbf{R}$. Define

$$\gamma^+(a_1, a_2) = \begin{cases} \gamma(a_1, a_2) \text{ if } a_1, a_2 \in \mathcal{A} \text{ or } a_1, a_2 \notin \mathcal{A} \\ \\ -\gamma(a_1, a_2) \text{ if } a_1 \in \mathcal{A}; a_2 \notin \mathcal{A} \text{ or } a_2 \in \mathcal{A}; a_1 \notin \mathcal{A}. \end{cases}$$

Symmetry of γ^+ is clear from inspection and additivity follows from the definition. To show γ^+ is self-positive, suppose there were $a \in \mathcal{A}^+$ such that $\gamma(a, a) \leq 0$. Any $a \in \mathcal{A}^+$ can be expressed as the composition an element in $\mathcal{A}$ and an inverse of some element: $a = a_1 \circ (-a_2)$ for some $a_1, a_2 \in \mathcal{A}$. If γ^+ were not self-positive, we would have $\gamma(a, a) \leq 0$ for some $a \neq a_0$, but then this would imply

$$\gamma^+(a_1 \circ (-a_2), a_1 \circ (-a_2)) = \gamma(a_1, a_1) + \gamma(a_2, a_2) - 2\gamma(a_1, a_2) \leq 0,$$

a violation of the positive definiteness of γ on $\mathcal{A}$ itself. This verifies the extension to $\mathcal{A}^+$.

Now consider $\mathcal{A}^*$. Let $\{\bar{a}_i\}_{i=1}^{d(\mathcal{A})}$ be a basis for $\mathcal{A}^*$ contained in the image of the embedding $f(\mathcal{A}) \subset \mathcal{A}^*$. Then for all $a_1 = \sum_{i=1}^{d(\mathcal{A})} q_i(a_1)\bar{a}_i \in \mathcal{A}^*$ and $a_2 = \sum_{i=1}^{d(\mathcal{A})} q_i(a_2)\bar{a}_i \in \mathcal{A}^*$, define

$$\gamma^*(a_1, a_2) = \sum_{i,j=1}^{d(\mathcal{A})} q_i(a_1) \cdot q_j(a_2) \cdot \gamma(f^{-1}(\bar{a}_i), f^{-1}(\bar{a}_j)).$$

We show γ^* is indeed an additive agreement function. Symmetry is clear from inspection and additivity follows from the definition. To show self-positivity, we note that any violation of self-positivity would imply a corresponding violation of self-positivity of γ^+: if there were $a = \sum_{i=1}^{d(\mathcal{A})} q_i(a)\bar{a}_i \in \mathcal{A}^*$, $a \neq a_0$ such that $\gamma^*(a, a) \leq 0$, then, denoting $C \equiv \text{g. c. d.}_{i=1,\ldots,d(\mathcal{A})}\{q_i(a)\}$,[17] it follows

$$\gamma^*(Ca, Ca) = \sum_{i,j=1}^{d(\mathcal{A})} \underbrace{Cq_i(a) \cdot Cq_j(a)}_{\in \mathbf{Z}} \cdot \gamma^*(\bar{a}_i, \bar{a}_j) = \gamma^+(Ca, Ca) \leq 0.$$

Therefore γ^* must be self-positive. Finally we consider uniqueness. Suppose that γ_1^* and γ_2^* are two extensions of γ to $\mathcal{A}^*$. If γ_1^* and γ_2^* are distinct, then it must be that $\gamma_1^*(\bar{a}_1, \bar{a}_2) \neq \gamma_2^*(\bar{a}_1, \bar{a}_2)$ for

[17] Recall that even if $d(\mathcal{A}) = \infty$, $q_i = 0$ for all but finitely many i, so C remains well defined.

some basis elements $\bar{a}_1, \bar{a}_2 \in \mathcal{A}^*$, but this contradicts the fact that, as they are extensions, it must be that $\gamma_1^*(\bar{a}_1, \bar{a}_2) = \gamma(f^{-1}(\bar{a}_1), f^{-1}(\bar{a}_2)) = \gamma_2^*(\bar{a}_1, \bar{a}_2)$, where $f : \mathcal{A} \longrightarrow \mathcal{A}^*$. Conclude that $\gamma_1^* = \gamma_2^*$. This proves statement (a).

To show statement (b), suppose that $a_1, a_2 \in \mathcal{A}$ are such that $\gamma(a_1, a_3) = \gamma(a_2, a_3)$ for all $a_3 \in \mathcal{A}$. Then by additivity

$$\begin{aligned}\gamma(a_2 - a_1, a_2 - a_1) &= \gamma(a_2, a_2) + \gamma(a_1, a_1) - 2\gamma(a_1, a_2) \\ &= \gamma(a_1, a_2) + \gamma(a_1, a_2) - 2\gamma(a_1, a_2) = 0,\end{aligned}$$

so by self-positivity $a_2 - a_1 = a_0$, and hence $a_1 = a_2$. This finishes the proof of the lemma.

Now, suppose $(\mathcal{A}, \mathcal{B})$ is a Virtual Bayesian. By Theorem 1, it satisfies (A.1,2,3), and by the proof of Theorem 1 there exists an essential embedding $f : \mathcal{A} \longrightarrow \mathcal{A}^*$ for some countable dimensional rational vector space $\mathcal{A}^*$. Let $(\bar{a}_i)_{i=1}^{d(\mathcal{A})}$ be a basis for $\mathcal{A}^*$, so each $a \in \mathcal{A}^*$ is of the form $a = \sum_{i=1}^{d(\mathcal{A})} q_i(a)\bar{a}_i$. We can easily define an additive agreement function on $\mathcal{A}^*$ by treating the basis elements as mutually orthogonal. Let

$$\gamma^*(a_1, a_2) \equiv \sum_{i=1}^{d(\mathcal{A})} q_i(a_1) \cdot q_i(a_2).$$

The symmetry condition $\gamma^*(a_1, a_2) = \gamma^*(a_2, a_1)$ holds by inspection. Additivity follows as

$$\begin{aligned}\gamma^*(a_1 \circ a_2, a) &= \sum_{i=1}^{d(\mathcal{A})} q_i(a_1 \circ a_2) \cdot q_i(a) \\ &= \sum_{i=1}^{d(\mathcal{A})} (q_i(a_1) + q_i(a_2)) \cdot q_i(a) \\ &= \sum_{i=1}^{d(\mathcal{A})} q_i(a_1) \cdot q_i(a) + \sum_{i=1}^{d(\mathcal{A})} q_i(a_2) \cdot q_i(a) = \gamma^*(a_1, a) + \gamma^*(a_2, a).\end{aligned}$$

As $\mathcal{A}^*$ is invertible, to demonstrate positive-definiteness it suffices to verify that every argument $a \in \mathcal{A}^*, a \neq a_0$, has positive self-agreement. As $a \neq a_0$, there must exist some i^* such that $q_{i^*}(a) \neq 0$. Then

$$\gamma^*(a, a) = \sum_{i=1}^{d(\mathcal{A})} q_i(a) \cdot q_i(a) \geq q_{i^*}(a)^2 > 0.$$

This proves that γ^* is an additive agreement function. Let γ be the restriction of γ^* to $\mathcal{A}$. By the lemma above, γ is an additive agreement function.

Finally we complete the other direction of the proof. Let $(\mathcal{A}, \mathcal{B})$ be a learning rule satisfying (A.1,2,3) and equipped with the additive agreement function γ. It suffices to establish that $\mathcal{A}$ also satisfies (A.4,5,6) from Theorem 1. That $\mathcal{A}$ satisfies (A.6) is straightforward:

$$\gamma(a^k, a) = k\gamma(a, a) \neq \gamma(a, a) > 0$$

for all $k > 1$, so by the fact that γ identifies $\mathcal{A}$ (by the lemma above) it follows $a^k \neq a$. Next, $\mathcal{A}$ satisfies (A.5) as a simple consequence of the additivity of γ and the fact that $\mathbf{R}$ is commutative: for all $a_1, a_2, a \in \mathcal{A}$,

$$\gamma(a_1 \circ a_2, a) = \gamma(a_1, a) + \gamma(a_2, a) = \gamma(a_2, a) + \gamma(a_1, a) = \gamma(a_2 \circ a_1, a),$$

so by identification $a_1 \circ a_2 = a_2 \circ a_1$. Lastly, we verify (A.4) by first establishing that $\mathcal{A}$ satisfies the cancellation property. Suppose $a_1 \circ a_3 = a_2 \circ a_3$ for some $a_1, a_2, a_3 \in \mathcal{A}$, and let $a \in \mathcal{A}$ be any other argument. Then

$$\gamma(a_1, a) + \gamma(a_3, a) = \gamma(a_1 \circ a_3, a) = \gamma(a_2 \circ a_3, a) = \gamma(a_2, a) + \gamma(a_3, a)$$

by additivity, which shows $\gamma(a_1, a) = \gamma(a_2, a)$, and therefore by identification it follows $a_1 = a_2$, verifying the cancellation property. Finally we establish that the cancellation property implies (A.4). Suppose $a(b_1) = a(b_2)$ for some $a \in \mathcal{A}$ and $b_1, b_2 \in \mathcal{B}$. By $\mathcal{A}$ self-recording under the bijection $g : \mathcal{A} \longrightarrow \mathcal{B}$, it follows $b_1 = g(a_1)$ and $b_2 = g(a_2)$ for some $a_1, a_2 \in \mathcal{A}$, and therefore

$$g(a_1 \circ a) = a(g(a_1)) = a(b_1) = a(b_2) = a(g(a_2)) = g(a_2 \circ a).$$

Thus $a_1 \circ a = a_2 \circ a$, which by the cancellation property implies $a_1 = a_2$, and therefore $b_1 = g(a_1) = g(a_2) = b_2$. This completes the proof. $\circ$

Proof of Theorem 3.

We start with statement (a). Let γ^* be the unique extension of γ to $\mathcal{A}^*$ whose existence is guaranteed by the lemma in the proof of Theorem 2, and let $\{\bar{a}_i\}_{i=1}^{d(\mathcal{A})}$ be a countable basis for $\mathcal{A}^*$. Below is an iterative procedure for defining an embedding the basis of $\mathcal{A}^*$ into $\mathbf{R}^n$, $f : \{\bar{a}_i\}_{i=1}^{d(\mathcal{A})} \longrightarrow \mathbf{R}^n$, such that $\langle f(\bar{a}_i), f(\bar{a}_j)\rangle = \gamma^*(\bar{a}_i, \bar{a}_j)$ for all $i, j = 1, \ldots, d(\mathcal{A})$. This will then serve as the foundation for embedding the entire space $\mathcal{A} \longrightarrow \mathcal{A}^* \longrightarrow \mathbf{R}^n$.

The first step is to iteratively construct a maximal 'linearly independent' subset $\{\bar{a}_i\}_{\mathcal{I}}^* \subset \{\bar{a}_i\}_{i=1}^{d(\mathcal{A})}$. How this relates to linear independence in the true sense of the word will be made clear as the proof progresses. To define the set of 'independent' indices $\mathcal{I}^*$ we first iterate through all of $i = 1, \ldots, d(\mathcal{A})$ and at each step define $\mathcal{I}_i^*$ to be the provisional set of independent indices *through* index i. At each step, define

$$\Sigma_{i,i} \equiv \Big(\gamma(\bar{a}_j, \bar{a}_k)\Big)_{j,k\in\mathcal{I}_i^*}$$

to be the real-valued $|\mathcal{I}_i^*| \times |\mathcal{I}_i^*|$ matrix of the pairwise agreement values between the elements of $\mathcal{I}_i^*$. To start, let $\mathcal{I}_1^* = \{1\}$. For $i > 1$, first assume that $\Sigma_{i-1,i-1}$ has been shown to be invertible. Now define $\Sigma_{i,i-1} \equiv \Big(\gamma^*(\bar{a}_i, \bar{a}_j)\Big)_{j\in\mathcal{I}_{i-1}^*}$ to be the $|\mathcal{I}_{i-1}^*| \times 1$ dimensional vector of agreement values between $\bar{a}_i$ and the elements of $\mathcal{I}_{i-1}^*$. Informally, we can think of the value of $\Sigma'_{i,i-1}\Sigma_{i-1,i-1}^{-1}\Sigma_{i,i-1}$ as the square of the magnitude of the 'projection' of $\bar{a}_i$ onto the span of the elements of $\mathcal{I}_{i-1}^*$. Such a projected value may not exist in $\mathcal{A}^*$ itself, but the γ^* values tell us where it ought to lie were $\mathcal{A}^*$ embedded in $\mathbf{R}^n$, and therefore $\bar{a}_i$ is considered independent from the elements $\mathcal{I}_{i-1}^*$ only if the magnitude of $\bar{a}_i$ strictly exceeds the magnitude of its 'projection.' Formally, if

$$\gamma(\bar{a}_i, \bar{a}_i) - \Sigma'_{i,i-1}\Sigma_{i-1,i-1}^{-1}\Sigma_{i,i-1} > 0,$$

then we set $\mathcal{I}_i^* = \mathcal{I}_{i-1}^* \cup \{i\}$, and otherwise we set $\mathcal{I}_i^* = \mathcal{I}_{i-1}^*$. Now we show that the assumption of $\Sigma_{i,i}$ invertible was indeed justified. Note that $\Sigma_{1,1}$ is invertible as $\bar{a}_1 \neq a_0$, so by self-positivity, $\gamma^*(\bar{a}_1, \bar{a}_1) > 0$. For $i > 1$, note that we can express $\Sigma_{i,i}$ as a block-symmetric matrix,

$$\Sigma_{i,i} = \begin{pmatrix} \Sigma_{i-1,i-1} & \Sigma_{i,i-1} \\ \Sigma'_{i,i-1} & \gamma^*(\bar{a}_i, \bar{a}_i) \end{pmatrix},$$

and as such its determinant is given by

$$|\Sigma_{i,i}| = |\Sigma_{i-1,i-1}| \underbrace{(\gamma^*(\bar{a}_i, \bar{a}_i) - \Sigma'_{i,i-1}\Sigma^{-1}_{i-1,i-1}\Sigma_{i,i-1})}_{>0}.$$

Thus $\Sigma_{i,i}$ has positive determinant, and is therefore invertible, so long as $\Sigma_{i-1,i-1}$ is. Conclude that $\Sigma_{i,i}$ is invertible for all $i \geq 1$. Finally, having defined $\mathcal{I}^*_i$ for all $i \geq 1$, set $\mathcal{I}^* \equiv \bigcup_{i=1}^{d(\mathcal{A})} \mathcal{I}^*_i$ and denote $n = |\mathcal{I}^*|$.

Now we embed the 'linearly independent' subset $\{\bar{a}_i\}_{i \in \mathcal{I}^*}$ into $\mathbf{R}^n$. With slight abuse of notation, relabel $\mathcal{I}^* = \{1, \dots, n\}$. Let the standard orthonormal basis vectors of $\mathbf{R}^n$ be denoted $e_1, \dots e_n$, and denote $f : \{\bar{a}_i\}_{i=1}^n \longrightarrow \mathbf{R}^n$ by $\bar{a}_i \longmapsto \sum_{k=1}^n C_{i,k} e_k$, where the values of $C_{i,k}$ are defined as follows. For $i = 1$, set $C_{1,1} = \gamma(\bar{a}_1, \bar{a}_1)^{1/2}$ and $C_{1,k} = 0$ for $k > 1$. For $i > 1$, extending γ to the standard inner product requires

$$\langle f(\bar{a}_i), f(\bar{a}_j) \rangle = \sum_{k=1}^{i} C_{i,k} C_{j,k} = \gamma(\bar{a}_i, \bar{a}_j)$$

for all $j \leq i - 1$. Thus

$$\begin{pmatrix} C_{i,1} \\ \vdots \\ C_{i,i-1} \end{pmatrix} = \underbrace{\begin{pmatrix} C_{1,1} & \dots & C_{1,i-1} \\ \vdots & \ddots & \vdots \\ C_{i-1,1} & \dots & C_{i-1,i-1} \end{pmatrix}^{-1}}_{\mathbf{C}^{-1}_{i-1}} \begin{pmatrix} \gamma(\bar{a}_i, \bar{a}_1) \\ \vdots \\ \gamma(\bar{a}_i, \bar{a}_{i-1}) \end{pmatrix}.$$

(Note that by construction $\mathbf{C}_{i-1}$ is an upper triangular matrix with strictly positive diagonal entries; it thus has positive determinant and is indeed invertible.) We define $C_{i,i}$ so that the magnitude of $\bar{a}_i$ is preserved under f,

$$C_{i,i} = \sqrt{\gamma(\bar{a}_i, \bar{a}_i) - \Sigma'_{i,i-1}\Sigma^{-1}_{i-1,i-1}\Sigma_{i,i-1}},$$

and set $C_{i,k} = 0$ for $k > i$. By constructing f in this manner, we have guaranteed γ^* is preserved on $\mathcal{I}^*$.

Now we show that f can extend γ^* on all of $i = 1, \dots, d(\mathcal{A})$. For any non independent basis element $\bar{a}_i$, let $\Sigma_{i,i-1}$ and $\Sigma_{i-1,i-1}$ be defined as in the construction of $\mathcal{I}^*$. Now let

$$\begin{pmatrix} c_{i,1} & \dots & c_{i,|\mathcal{I}^*_{i-1,i-1}|} \end{pmatrix}' = \Sigma^{-1}_{i-1,i-1}\Sigma_{i,i-1},$$

and define $f(\bar{a}_i) = \sum_{j \in \mathcal{I}^*_{i-1}} c_{ij} f(\bar{a}_j)$. We show that for any arbitrary $a' \in \mathcal{A}^*$ it must be the case that $\gamma^*(\bar{a}_i, a') = \sum_{j \in \mathcal{I}^*_{i-1}} c_{ij} \gamma^*(\bar{a}_j, a')$, which verifies that f extends γ^* on the entire basis. To do this, let $(q_j)_{j \in \mathcal{I}^*_{i-1}} \in \mathbf{Q}$ be any arbitrary set of rational coefficients. Then, by the Cauchy-Schwartz inequality (which is proved for this setting as a lemma at the very end of the proof of the theorem):

$$\begin{aligned}\gamma^* \left(\bar{a}_i - \sum_{j \in \mathcal{I}^*_{i-1}} q_j \bar{a}_j, a' \right)^2 &\leq \gamma^*(a', a') \gamma^* \left(\bar{a}_i - \sum_{j \in \mathcal{I}^*_{i-1}} q_j \bar{a}_j, \bar{a}_i - \sum_{j \in \mathcal{I}^*_{i-1}} q_j \bar{a}_j \right) \\ &= \gamma^*(a', a') \underbrace{\left(\gamma^*(\bar{a}_i, \bar{a}_i) - 2 \sum_{j \in \mathcal{I}^*_{i-1}} q_j \gamma^*(\bar{a}_i, \bar{a}_j) + \sum_{j,k \in \mathcal{I}^*_{i-1}} q_j q_k \gamma^*(\bar{a}_j, \bar{a}_k) \right)}_{\delta(q)}.\end{aligned}$$

Note that as $q_j \longrightarrow c_{ij}$ for all $j \in \mathcal{I}^*_{i-1}$,

$$\delta(q) \longrightarrow \gamma^*(\bar{a}_i, \bar{a}_i) - 2\Sigma'_{i,i-1} \Sigma^{-1}_{i-1,i-1} \Sigma_{i,i-1} + \Sigma'_{i,i-1} \Sigma^{-1}_{i-1,i-1} \Sigma_{i-1,i-1} \Sigma^{-1}_{i-1,i-1} \Sigma_{i,i-1} = 0.$$

Thus

$$\begin{aligned}\lim_{q_j \longrightarrow c_{ij}} \left(\gamma^*(\bar{a}_i, a') - \sum_{j \in \mathcal{I}^*_{i-1}} q_j \gamma^*(\bar{a}_j, a') \right)^2 &= \lim_{q_j \longrightarrow c_{ij}} \gamma^* \left(\bar{a}_i - \sum_{j \in \mathcal{I}^*_{i-1}} q_j \bar{a}_j, a' \right)^2 \\ &\leq \lim_{q_j \longrightarrow c_{ij}} \gamma^*(a', a') \delta(q) = 0,\end{aligned}$$

so $\gamma^*(\bar{a}_i, a') = \sum_{j \in \mathcal{I}^*_{i-1}} c_{ij} \gamma^*(\bar{a}_j, a')$, verifying f extends γ to all of $\{\bar{a}_i\}_{i=1,\ldots,d(\mathcal{A})}$.

Finally, setting $f(a) = \sum_{i=1}^{d(\mathcal{A})} q_i(a) f(\bar{a}_i)$ allows for γ to be extended for all of $\mathcal{A}^*$. The restriction of f to $\mathcal{A}$ constitutes the essential embedding described in the theorem statement. This proves statement (a) except for the uniqueness claim, which is handled below.

Now to show statement (b), suppose f is an essential embedding $\mathcal{A} \longrightarrow \mathbf{R}^n$. Then $\gamma(a_1, a_2) = \langle f(a_1), f(a_2) \rangle$ is an additive agreement function on $\mathcal{A}$. (All three properties easily follow from $\langle \cdot, \cdot \rangle$ being an inner product on $\mathbf{R}^n$). Now we establish the uniqueness claims of (a) and (b). Suppose $f_1 : \mathcal{A} \longrightarrow \mathbf{R}^n$ and $f_2 : \mathcal{A} \longrightarrow \mathbf{R}^n$ both extend the same γ on $\mathcal{A}$ to $\langle \cdot, \cdot \rangle$ on $\mathbf{R}^n$. As f_1 and f_2 are essential, there exists bases $\{e^1_i\}_{i=1}^n$ and $\{e^2_i\}_{i=1}^n$ for $\mathbf{R}^n$ contained in the images of f_1 and f_2. Define $g : \mathbf{R}^n \longrightarrow \mathbf{R}^n$ by

$$g : \sum_{i=1}^n x_i e^1_i \longmapsto \sum_{i=1}^n x_i f_2(f_1^{-1}(e^1_i)).$$

As f_1 and f_2 both preserve γ, it follows that $\langle x, y\rangle = \langle g(x), g(y)\rangle$ for all $x, y \in \mathbf{R}^n$. That is, g is an orthogonal transformation. Thus the embedding constructed in the proof of part (a) is unique up to orthogonal transformation, and the γ constructed in the proof of part (b) is unique.

Finally we establish statement (c). Let $\mathcal{A}$ be any Virtual Bayesian learning rule of algebraic dimension $d(\mathcal{A})$. Let $f^{d(\mathcal{A})} : \mathcal{A}^* \cong \mathbf{Q}^{d(\mathcal{A})} \longrightarrow \mathbf{R}^{d(\mathcal{A})}$ be the inclusion mapping from the $d(\mathcal{A})$ dimensional rational vector space to the $d(\mathcal{A})$ dimensional real vector space. Without loss of generality assume that the standard basis of $\mathbf{R}^{d(\mathcal{A})}$ is contained in $f(\mathcal{A}^*)$. This is an essential embedding. Now, to construct an essential embedding into $\mathbf{R}^n$ for $n < d(\mathcal{A})$, first let $\{x_i\}_{i=1}^{d(\mathcal{A})}$ be the set of 'mutually irrational' real numbers as was used in the proof of Theorem 1. Then set $f^n : \mathcal{A}^* \longrightarrow \mathbf{R}^n$ by

$$f^n\left(\sum_{i=1}^{d(\mathcal{A})} q_i(a)\bar{a}_i\right) = \sum_{i=1}^{n} q_i(a) f^{d(\mathcal{A})}(\bar{a}_i) + \sum_{i=n+1}^{d(\mathcal{A})} q_i(a)x_i.$$

This function modifies the inclusion mapping $f^{d(\mathcal{A})}$ by collapsing all dimensions higher than n onto the first dimension while still preserving additivity. As the dimensionality of $\mathbf{R}^{d(\mathcal{A})}$ is reduced in the process, $f^{d(\mathcal{A})}$ being essential guarantees that f^n is essential as well. This completes the proof.

Lemma. *(Cauchy-Schwartz for γ^* on $\mathcal{A}^*$) $\gamma^*(a_1, a_2)^2 \leq \gamma^*(a_1, a_1) \cdot \gamma^*(a_2, a_2)$ for all $a_1, a_2 \in \mathcal{A}^*$.*

Proof. Consider that by positive-definiteness it follows for any $a_1, a_2 \in \mathcal{A}^*$ and $q_1, q_2 \in \mathbf{Q}$ that

$$\gamma^*(q_1a_1 - q_2a_2, q_1a_1 - q_2a_2) = q_1^2\gamma^*(a_1, a_1) + q_2^2\gamma^*(a_2, a_2) - 2q_1q_2\gamma^*(a_1, a_2) \geq 0.$$

Let $(q_1^k), (q_2^k) \in \mathbf{Q}$ be sequences with $q_1^k \longrightarrow \gamma^*(a_1, a_1)^{-1/2}$ and $q_2^k \longrightarrow \gamma^*(a_2, a_2)^{-1/2}$ as $k \longrightarrow \infty$. Then, as

$$\gamma^*(a_1, a_2) \leq \frac{1}{q_1^k q_2^k} \cdot \frac{1}{2}\left((q_1^k)^2\gamma^*(a_1, a_1) + (q_2^k)^2\gamma^*(a_2, a_2)\right)$$

for all $k = 1, 2, \ldots$, it follows

$$\begin{aligned}\gamma^*(a_1, a_2) &\leq \left(\frac{1}{\gamma^*(a_1, a_1)^{-1/2} \cdot \gamma^*(a_2, a_2)^{-1/2}}\right) \cdot \frac{1}{2}\left(\frac{\gamma^*(a_1, a_1)}{\gamma^*(a_1, a_1)} + \frac{\gamma^*(a_2, a_2)}{\gamma^*(a_2, a_2)}\right)\\ &= \sqrt{\gamma^*(a_1, a_1) \cdot \gamma^*(a_2, a_2)},\end{aligned}$$

and squaring both sides delivers the Cauchy-Schwartz inequality. ○

Proof of Proposition 3. The proof proceeds in two steps. We first construct a homeomorphism which warps (almost all of) the unit sphere into a region of the unit cube, after which the connectedness and dimensionality claims in the proposition statement follow straightforwardly.

To deform the sphere, first fix an arbitrary numeraire state $x_0 \in \mathcal{X}$, enumerate the remaining states $x_1, \ldots, x_n$, let $(\bar{a}_i)_{i=1}^n$ be a basis for $\mathcal{A}$, and let $l_i \in \mathbf{R}^n$ be the log likelihood-ratio with respect to x_0 for argument $\bar{a}_i$. For each $y \in S^{n-1}$, where $y = \sum_{i=1}^{d(\mathcal{A})} c_i \bar{a}_i$, define

$$\hat{l}(y) = \sum_{i=1}^{d(\mathcal{A})} c_i l_i = Ly,$$

to be the implied log likelihood-ratio at y. By the full-dimensionality assumption, the set (l_i) is linearly independent, so L is an invertible matrix. Define $S(\chi) = \{y \in S \mid y_i > 0, y_i \geq y_j \forall j = 1, \ldots, n \iff x_i \in \chi\}$ to be the subset of directions in which the maximum log likelihood-ratio states are χ. By inspection, any argument in this region reaches χ. This construction applies only for $x_0 \notin \chi$, but – as is also done in the rest of the proof – to consider a region corresponding to χ with $x_0 \in \chi$, one need only choose an alternative numeraire state.

Note that $\hat{l}(y) = (\hat{l}_j(y))_{j=1}^{n-1}$ is a $n-1$ dimensional vector, and denote $\hat{l}^{\mathsf{max}}(y) = \max_{j=1,\ldots,n-1} l_j(y)$ to be the maximum of the different non-numeraire states' log likelihood-ratios. Now let $\mathcal{Y} \subset S^{n-1}$ denote the set of y for which $\hat{l}^{\mathsf{max}}(y) > 0$, the union of all regions $S(\chi)$ with $x_0 \notin \chi$. Define the transformation

$$\psi : y \longmapsto l(y) + \frac{1 - l^{\mathsf{max}}(y)}{1 + l^{\mathsf{max}}(y)} \cdot (l(y) + 1).$$

This is the 'projection' of from the point $(-1, -1, \ldots, -1)$ through $\mathcal{Y}$ onto the cube $[-1, 1]^n$. Let $\mathcal{Z}$ denote $\psi(\mathcal{Y})$. As ψ is continuous and admits the continuous inverse

$$\psi^{-1} : z \longmapsto L^{-1} \left(\frac{z - k(z)}{1 + k(z)} \right); \quad k(z) = (n-1)^{-1} \left(1 + \sum_{i=1}^n z_i + \sqrt{\left(1 + \sum_{i=1}^n z_i\right)^2 - (n-1)\left(-1 + \sum_{i=1}^n z_i^2\right)} \right),$$

it follows that $\mathcal{Y}$ and $\mathcal{Z}$ are homeomorphic.

The usefulness of the transformation is that ψ carries each region $S(\chi)$ to a single hyperface of the cube:

$$\psi(S(\chi)) = \{z \in \mathcal{Z} \mid z_i = 1 \iff x_i \in \chi, \ i = 1, \ldots, n\}.$$

Note further that each $\psi(S(\chi))$, as the intersection of a cube hyperface and – as ψ is a projection of $\mathcal{Y}$ from $(-1,-1,\ldots,-1)$ – a convex cone. We can now quite easily prove statements (a-c) about the $\psi-$transformed regions lying in $\mathcal{Z}$. Then, as ψ^{-1} is a homeomorphism, this implies that (a-c) also hold for the un-transformed regions lying in $\mathcal{Y}$.

First we show $\psi(S(\chi))$, $\chi \neq \{x_0\}$, is itself path-connected. Let $z_1, z_2 \in \psi(S(\chi))$ and consider the straight-line path $[0,1] \longrightarrow \mathcal{Z}$, $\alpha \longmapsto (1-\alpha)z_1 + \alpha z_2$. For any i with $x_i \in \chi$, $z_{1,i} = z_{2,i} = 1$, so $((1-\alpha)z_1 + \alpha z_2)_i = (1-\alpha) + \alpha = 1$; for any i with $x_i \notin \chi$, $z_{1,i}, z_{2,i} < 1$, so $((1-\alpha)z_1 + \alpha z_2)_i < (1-\alpha)+\alpha = 1$. This shows the entire path is contained in $\psi(S(\chi))$, so the region is path-connected. Finally, to establish $S(\{x_0\})$ is path-connected, simply choose another numeraire state $x_0' \neq x_0$ and repeat the above argument.

Next, suppose $\chi_1 \subset \chi_2$, $x_0 \notin \chi_1$, are neighboring. Let $z_j \in \psi(S(\chi_j))$, $j = 1,2$, be defined by

$$z_{j,i} = \begin{cases} 1 \text{ if } x_i \in \chi_j \\ 1+\varepsilon \text{ otherwise,} \end{cases}$$

where $\varepsilon > 0$ is sufficiently small that $z_j \in \psi(S(\chi_j))$, $j = 1,2$. Then define the path $[0,1] \longrightarrow \mathcal{Z}$, $\alpha \longmapsto (1-\alpha)z_1 + \alpha z_2$ as before. For $\alpha < 1$, $((1-\alpha)z_1 + \alpha z_2)_i = 1$ if and only if $x_i \in \chi_1$; for $\alpha = 1$, $((1-\alpha)z_1 + \alpha z_2)_i = 1$ if and only if $x_i \in \chi_2$. This shows the region $\psi(S(\chi_1)) \cup \psi(S(\chi_2)) = \psi(S(\chi_1) \cup S(\chi_2))$ is path-connected. To cover the case of $x_0 \in \chi_1 \subset \chi_2$, simply choose another numeraire $x_0' \notin \chi_2$ and repeat the argument.

Now suppose χ_1 and χ_2 are not neighboring, so there exists $x_i \in \chi_1 \setminus \chi_2$ and $x_j \in \chi_2 \setminus \chi_1$. (For this portion the ψ transformation proves more hindrance than help, so we temporarily put it aside.) Let $f : [0,1] \longrightarrow \mathcal{Y}$ be any path from $y_1 \in S(\chi_1)$ to $y_2 \in S(\chi_2)$. By the choice of i and j it follows $f(0)_i - f(0)_j > 0$ and $f(1)_i - f(1)_j < 0$, so by the intermediate value theorem, there exists $\alpha \in (0,1)$ such that $f(\alpha)_i = f(\alpha)_j$, meaning $f(\alpha)$ lies neither in $S(\chi_1)$ nor $S(\chi_2)$. Thus $S(\chi_1) \cup S(\chi_2)$ is not path-connected. Conclude χ_1 and χ_2 are neighboring if and only if $S(\chi_1) \cup S(\chi_2)$ is path-connected.

Finally we establish the dimensionality of $\psi(S(\chi))$. For the sake of notation, let $\chi = \{x_1, \ldots, x_m\}$. Any $z \in \psi(S(\chi))$ is of the form $z = (1, \ldots, 1, z_{m+1}, \ldots, z_{m+(n-m)})$, where $z_k < 1$, $k = 1, \ldots, n-m$. Let $\bar{\varepsilon} > 0$ be small enough that the (square) ball

$$\mathcal{B}_{\varepsilon}(z) \equiv \left\{(1, \ldots, 1, z_{m+1} + \varepsilon_1, \ldots, z_{m+(n-m)} + \varepsilon_{n-m}) \mid |\varepsilon_k| < \bar{\varepsilon}, k = 1, \ldots, n-m\right\}$$

is contained in $\psi(S(\chi))$. The mapping

$$(1, \ldots, 1, z_{m+1} + \varepsilon_1, \ldots, z_{m+(n-m)} + \varepsilon_{n-m}) \longmapsto \left(\frac{\varepsilon_1}{\bar{\varepsilon}^2 - \varepsilon_1^2}, \ldots, \frac{\varepsilon_{n-m}}{\bar{\varepsilon}^2 - \varepsilon_{n-m}^2} \right)$$

is a homeomorphism $\mathcal{B}_\varepsilon(z) \longrightarrow \mathbf{R}^{n-m}$, so $\psi(S(\chi))$, and therefore $S(\chi)$ also, is a manifold of dimension $n - m = |\mathcal{X}| - |\chi| - 1$.

$\circ$