General properties of general Bayesian learning

Zalán Gyenis[*] Miklós Rédei[†]

August 22, 2015

Abstract

We investigate the general properties of general Bayesian learning, where "general Bayesian learning" means inferring a state from another that is regarded as evidence, and where the inference is conditionalizing the evidence using the conditional expectation determined by a reference probability measure representing the background subjective degrees of belief of a Bayesian Agent performing the inference. States are linear functionals that encode probability measures by assigning expectation values to random variables via integrating them with respect to the probability measure. If a state can be learned from another this way, then it is said to be Bayes accessible from the evidence. It is shown that the Bayes accessibility relation is reflexive, antisymmetric and non-transitive. If every state is Bayes accessible from some other defined on the same set of random variables, then the set of states is called weakly Bayes connected. It is shown that the set of states is not weakly Bayes connected if the probability space is standard. The set of states is called weakly Bayes connectable if, given any state, the probability space can be extended in such a way that the given state becomes Bayes accessible from some other state in the extended space. It is shown that probability spaces are weakly Bayes connectable. Since conditioning using the theory of conditional expectations includes both Bayes' rule and Jeffrey conditionalization as special cases, the results presented generalize substantially some results obtained earlier for Jeffrey conditionalization.

1 Review of main results

In this paper we investigate the general properties of general Bayesian learning. By "general Bayesian learning" we mean inferring a probability measure from another that is regarded as evidence, and where the inference is conditionalizing the probability measure representing the evidence using the conditional expectation determined by a reference probability measure that is interpreted as representing the background subjective degrees of belief of a Bayesian Agent performing the inference.

The investigation is motivated by the observation that the properties of Bayesian learning we wish to determine do not seem to have been analyzed in the literature on Bayesianism on the level of generality we aim at here. (For monographic works on Bayesianism we refer to [22], [3], [44]; for papers discussing basic aspects of Bayesianism see [21], [19], [20]; the recent paper by Weisberg [43] provides a compact review of Bayesianism). In particular, in this paper we take the position that the proper general technical device to perform Bayesian conditioning is the theory of conditional

[*] MTA Rényi Institute of Mathematics, Budapest, Hungary, gyz@renyi.hu

[†] Department of Philosophy, Logic and Scientific Method, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK, m.redei@lse.ac.uk

expectations. The concept of conditional expectation was introduced into probability theory by Kolmogorov in 1933 together with his axiomatization of probability theory, which has made probability theory part of measure theory [27] (Doob [8] puts Kolmogorov's work into historical context). Since Kolmogorov's work, conditioning using the theory of conditional expectations has become standard in mathematics [9], [12], [1], [34], [2], [32]. Both the elementary Bayes' rule (sometimes called "strict conditionalization") and Jeffrey conditionalization (also called "probability kinematics" [23]) can be recovered as special cases of conditioning using the theory of conditional expectations; although, somewhat surprisingly, the fact that Jeffrey conditionalization is indeed a special case of conditioning via conditional expectations does not seem to be well known: Jeffrey does not refer to the theory of conditional expectations when introducing his rule of conditionalization in [23], nor do standard mathematical works on probability theory [12], [1], [34], [2], [32] mention Jeffrey conditionalization when discussing the concept of conditional expectation.

Proper handling of conditioning via conditional expectation requires one to go beyond the framework of additive measures on Boolean algebras and forces one to work with positive, normalized, linear functionals (called "states") that assign finite expectation values to random variables via integrating them with respect to probability measures on the Boolean algebra. Viewed from this more general perspective, conditioning can be regarded as a map in the state space that takes a state as input and yields another, the conditioned state, as output. Conditioning is thus a map in the dual space of the function space consisting of integrable random variables. *Bayesian* conditioning is distinguished among the logically possible conditioning maps in the dual space by the fact that it is the dual of a specific map (a projection) on the function space representing the random variables that are integrable with respect to the background probability of the Bayesian Agent. This projection on the function space is the conditional expectation. Bayesian conditionalization based on the technique of conditional expectations as conditioning device thus defines a two-place relation in the state space of integrable random variables. We call this relation the "Bayes accessibility relation" (Definition 4.3). The interpretation of the Bayes accessibility relation is that if a state is Bayes accessible from another, then the Bayesian agent can infer this state from the evidence represented by the other state, where the inference is a Bayesian upgrading using the technique of conditional expectation determined by the Agent's background probability. Characterizing the Bayes accessibility relation amounts then to characterizing Bayesian learning in this general setting. This is what the present paper does.

The first result of the paper is that the Bayes accessibility relation is antisymmetric (Proposition 5.1). (The Bayes accessibility relation is trivially reflexive.) Antisymmetry of the Bayes accessibility relation entails that a state space is *not* strongly Bayes connected: It is *not* true that any two states are Bayes accessible from each other. In other words, Bayesian learning has a certain directedness built into it: if a state can be learned from another that represents evidence, then the converse is not true, Bayesian learning in the reversed direction is not possible (assuming that the two states are different). This seems an intuitively attractive feature of Bayesian learning: what one can learn from evidence cannot serve as evidence to learn the evidence itself.

We then ask the question of whether state spaces are *weakly* Bayes connected: is it true that *every* state is Bayes accessible from *some* other state (Definition 6.1)? Weak Bayes connectedness means that for every probability measure there exists some evidence from which the Bayesian Agent can learn that probability by conditionalization with respect to his fixed background degree of belief. Failure of weak Bayes connectedness means that, given the background measure of the Bayesian Agent, there exist states (probability measures) that the Agent cannot learn via Bayesian upgrading no matter what evidence formulated in the given state space he is presented with. Thus weak Bayes connectedness of the state space would be a sign of strength of Bayesian learning – failure of weak connectedness sets a limit to Bayesian learning in the given context. We give a characterization of weak Bayes connectedness of state spaces (Proposition 6.2). This result is used then to show that state spaces of *standard* probability spaces are *not* weakly connected (Propositions 6.7 and 6.8). Standard probability spaces include essentially all the probability spaces that occur in applications of probability theory; in particular, probability spaces with a finite number of random events, and probability theories in which probability is given by a density function with respect to the Lebesgue measure, are standard. In fact, we prove more: in case of a standard probability space there exist an uncountably infinite number of probability measures that are inaccessible for the Bayesian Agent (Propositions 6.10 and 6.11). Note that since conditioning now is with respect to conditional expectations, not via the simple Bayes' rule, the existence of Bayes inaccessible states has nothing to do with the well known fact that a measure which is obtained from another via the simple Bayes' rule will have to take zero value whenever the prior measure takes on value zero and that therefore a lot of measures (for instance all faithful probability measures) cannot be obtained as the result of conditionalizing another measure via the simple Bayes' rule. (A probability measure is faithful if all non-zero events have non-zero probability. A faithful *conditional* probability measure appears in Example 7.2.)

Proposition 6.2, which characterizes weak Bayes connectedness, makes it possible to formulate a condition sufficient to entail that a state space is weakly Bayes connected (Proposition 6.5). Based on this latter condition we give an example of a probability space whose state space is weakly Bayes connected. The significance for Bayesian learning of the probability space of this weakly Bayes connected state space might be very limited however because it will be seen that the cardinality of the Boolean algebra of this weakly Bayes connected state space is much larger than that of the continuum. Thus, only Bayesian Agents capable of comprehending an enormous amount of propositions would be in the position to have degrees of belief in every proposition in such a large set. Whether the notion of Bayesian Agent should include Agents with such extraordinary mental skills, is questionable. In the typical situations when one deals with probabilistic modeling, the concept of a Bayesian Agent with more modest mental powers is sufficient. But in such contexts inaccessibility of certain states via Bayesian inference is the general rule.

Failure of weak Bayes connectedness leads to the question of whether state spaces are weakly Bayes *connectable*: Whether for every state (in particular for a state that is not Bayes accessible from any other state in the given probabilistic framework) there exists a richer probability theory into which the original can be embedded in such a way that the Bayes inaccessible state becomes Bayes accessible from *some* state in the richer framework. We show that state spaces are weakly Bayes connectable (Proposition 8.3). This result generalizes the ones obtained by Diaconis and Zabell for the simple Bayes' rule and for Jeffrey conditionalization [6] (also see [17]). Weak Bayes connectability of state spaces means that everything that can be formulated by the Bayesian Agent in terms of a given probability space, can *in principle* be learned by the Agent by Bayesian upgrading – provided that the Agent is allowed to have access to a rich enough pool of evidence. We will call this latter upgrading situation "Unlimited Evidence Upgrading" scenario, in contradistinction to the "Limited Evidence Upgrading" situation, in which the evidence available to the Agent is restricted to the set of all states on a given set of random variables. Thus under the Unlimited Evidence Upgrading conditions a Bayesian Agent has unlimited Bayesain learning capacity.

Weak Bayes connectability of state spaces raises the question of whether state spaces are *strongly* Bayes connectable: whether it is true that any state can be made Bayes accessible from any other by embedding both into a larger state space (Problem 8.6). This problem remains open.

We will also show that the Bayes accessibility relation is *not* transitive (Proposition 7.1). Non-transitivity of Bayes accessibility means that while the Bayesian Agent might be able to learn a state from another in several successive steps of upgrading, the Agent will not in general be able to cut short the learning process by replacing the chain of steps leading to learning a state by a single Bayesian learning move – failure of transitivity of Bayes accessibility means that "There is no Bayesian royal road to learning".

The proof of failure of transitivity of the Bayes accessibility relation reveals that the reason behind this feature is the non-commutativity of general Bayesian upgrading via conditional expectations: the result of upgrading more than once depends on the order of the upgradings. Non-commutativity of Jeffrey conditionalization has been known for long and has been analyzed in a number of papers [5], [13], [7], [16], [6], [39], [4], [41], [15], [42]. The result presented here shows that non-commutativity is a general feature of general Bayesian learning; it is a feature that is linked to the very essence of Bayesian upgrading via conditional expectations. Weisberg [41] (in harmony with others) diagnoses the source of failure of commutativity of upgrading via Jeffrey conditionalization in what is called "rigidity" of upgrading. Roughly put, rigidity is the feature that the conditioned state preserves the evidence. Referring to general, non-trivial results concerning characterization of conditional expectations we will show that Weisberg's diagnosis is deep and applies to general Bayesian learning as well. We will also argue however that failure of commutativity of upgrading via conditional expectations is philosophically not as problematic as it might appear. The standard interpretation of non-commutativity of upgrading using Jeffrey conditionalization is that it violates an important general norm of rationality (which we dub here the "Norm of Epistemic Commutativity"): the demand that conclusions drawn from some body of evidence should not depend on the order in which elements of that evidence are presented. (For recent articulations of this interpretation of failure of non-commutativity see [41], [15], [42]). We will argue that non-commutativity of the conditional expectations Bayesian upgrading is based on should not be interpreted as violation of the Norm of Epistemic Commutativity. We also will show that, under a proper understanding of the concept of evidence in a Bayesian upgrading, and under a technically explicit specification of the Norm of Epistemic Commutativity in terms of conditional expectations, general Bayesian learning (hence also Jeffrey conditionalization) does satisfy the Norm of Epistemic Commutativity in spite of the upgrading being non-commutative.

The structure of the paper is the following. Section 2 fixes notation and recalls some basic definitions and facts from the theory of conditional expectations. Section 3 defines conditional probability in terms of conditional expectations and shows how elementary Bayesian upgrading and Jeffrey conditionalization obtain as special cases of conditionalization via conditional expectation. Section 4 defines the Bayes accessibility relation, illustrates Bayes accessibility by an example that involves conditionalization that cannot be handled by either simple Bayes rule or Jeffrey conditionalization. Section 5 proves that the Bayes accessibility relation is antisymmetric and discusses failure of strong Bayes connectedness of state spaces. Section 6 analyses weak Bayes connectedness and proves that state spaces of standard probability spaces are not weakly Bayes connected. Section 7 proves failure of transitivity of Bayes accessibility and discusses both Weisberg's rigidity analysis and why the Norm of Epistemic Commutativity is in full harmony with Bayesian upgrading via conditional expectations. Section 8 proves weak Bayes connectability of state spaces. Section 9 summarizes the main points with some further comments.

2 Conditional expectations

We fix some notation that will be used throughout the paper. $(X, \mathcal{S}, p)$ denotes a probability measure space: X is the set of elementary events, $\mathcal{S}$ is a σ-algebra of some subsets of X, p is a probability measure on $\mathcal{S}$. Given $(X, \mathcal{S}, p)$, $\mathcal{L}^s(X, \mathcal{S}, p)$ denotes the set of $f\colon X \to \mathbb{R}$ measurable functions such that $|f|^s$ is p-integrable. Of special importance are the integrable ($s = 1$), the square-integrable ($s = 2$), and the (essentially) bounded functions, the latter corresponds, formally, to $s = \infty$. Since p is a bounded measure, we have (cf. [38], [35][p. 71])

$$\mathcal{L}^\infty(X, \mathcal{S}, p) \subset \mathcal{L}^2(X, \mathcal{S}, p) \subset \mathcal{L}^1(X, \mathcal{S}, p) \tag{1}$$

Identifying functions that are equal except on p-measure zero sets, one obtains the corresponding spaces $L^s(X, \mathcal{S}, p)$ consisting of equivalence classes of functions (notice the notational difference between $\mathcal{L}$ and L). In what follows, in harmony with the usual mathematical practice, we use the same letters f, g etc. to refer to both functions (elements of $\mathcal{L}^s(X, \mathcal{S}, p)$) and equivalence classes of functions (elements of $L^s(X, \mathcal{S}, p)$). The characteristic (indicator) functions χ_A of the sets $A \in \mathcal{S}$ are in $\mathcal{L}^s(X, \mathcal{S}, p)$ for all $A \in \mathcal{S}$.

The probability measure p extends from $\mathcal{S}$ to a linear functional ϕ_p on $\mathcal{L}^s(X, \mathcal{S}, p)$ by the integral:

$$\phi_p(f) \doteq \int_X f^s dp \qquad f \in \mathcal{L}^s(X, \mathcal{S}, p) \tag{2}$$

The value of ϕ_p on a characteristic function χ_A of $A \in \mathcal{S}$ is just the p-probability of A:

$$\phi_p(\chi_A) = \int_X \chi_A^s dp = \int_X \chi_A dp = \int_A dp = p(A) \tag{3}$$

The map $f \mapsto \|f\|_s \doteq \phi_p(|f|)$ defines a seminorm $\|\cdot\|_s$ on $\mathcal{L}^s(X, \mathcal{S}, p)$ (only a *semi*norm because in the function space $\mathcal{L}^s(X, \mathcal{S}, p)$ functions differing on p-probability zero sets are *not* identified). The linear functional ϕ_p is continuous in the seminorm $\|\cdot\|_s$. The seminorm $\|\cdot\|_s$ becomes a norm on $L^s(X, \mathcal{S}, p)$. The containment relation (1) is dense when $\mathcal{L}^s$ ($s = 1, 2, \infty$) are considered as normed spaces.

We denote by $L^1(X, \mathcal{S}, p)^\sharp$ the set of all linear functionals on $L^1(X, \mathcal{S}, p)$ that are positive $\phi(f) \geq 0$ if $f \geq 0$, and normalized $\phi(\mathbf{1}) = 1$, where $\mathbf{1}$ denotes the characteristic function χ_X of the whole set X. These functionals are called *states*, and these are the ones that define probability measures when restricted to the (characteristic functions of elements of the) σ-algebra $\mathcal{S}$ and which thus encode probability measures via the integral (2). Note that because of the requirement of boundedness, any $\phi \in L^1(X, \mathcal{S}, p)^\sharp$ is continuous in the $\|\cdot\|_1$-norm topology; thus the states are a proper subset of the dual space $L^1(X, \mathcal{S}, p)^*$, the latter containing *all* $\|\cdot\|_1$-continuous linear functionals.

The space of square-integrable random variables $L^2(X, \mathcal{S}, p)$ is a Hilbert space with respect to the scalar product $\langle\cdot,\cdot\rangle$ defined by

$$\langle f, g\rangle \doteq \int_X fg \; dp \qquad f, g \in L^2(X, \mathcal{S}, p) \tag{4}$$

For more details on the above notions (and other mathematical concepts related to L^s-spaces used here without definition) see the standard references for the measure theoretic probability theory [29], [1], [34], [2]. In particular, section 19 in [1] and Chapter 3 in [35] discuss further properties of the function spaces $L^s(X, \mathcal{S}, p)$.

The central concept that the modern mathematical theory of conditionalization is based on is the notion of conditional expectation:

Definition 2.1 ([1] p. 445)**.** Let $(X, \mathcal{S}, p)$ be a probability space, $\mathcal{A}$ be a σ-subalgebra of $\mathcal{S}$, and $p_{\mathcal{A}}$ be the restriction of p to $\mathcal{A}$. A map

$$\mathscr{E}(\cdot \mid \mathcal{A}) \colon \mathcal{L}^1(X, \mathcal{S}, p) \to \mathcal{L}^1(X, \mathcal{A}, p_{\mathcal{A}}) \tag{5}$$

is called an $\mathcal{A}$-conditional expectation from $\mathcal{L}^1(X, \mathcal{S}, p)$ to $\mathcal{L}^1(X, \mathcal{A}, p_{\mathcal{A}})$ if (i) and (ii) below hold:

(i) For all $f \in \mathcal{L}^1(X, \mathcal{S}, p)$, the $\mathscr{E}(f \mid \mathcal{A})$ is $\mathcal{A}$-measurable.

(ii) $\mathscr{E}(\cdot \mid \mathcal{A})$ preserves the integration on elements of $\mathcal{A}$:

$$\int_Z \mathscr{E}(f \mid \mathcal{A}) dp_{\mathcal{A}} = \int_Z f \; dp \qquad \forall Z \in \mathcal{A}. \tag{6}$$

The $\mathcal{A}$-measurability condition (i) should be thought of as a coarse-graining requirement: it entails that $\mathscr{E}(f \mid \mathcal{A})$ is constant on minimal elements (atoms) in $\mathcal{A}$ (atoms in $\mathcal{A}$ need not be atoms in $\mathcal{S}$). Condition (ii) is the general form of the theorem of total probability: it requires that from the conditional expectation one can recover the original expectation values (hence the original probability p). For further discussion of the interpretation of properties of the conditional expectation see [1].

It is not obvious that, given an $\mathcal{A}$, a conditional expectation $\mathscr{E}(\cdot \mid \mathcal{A})$ exists but the Radon-Nykodim theorem entails that it *always* does:

Proposition 2.2 ([1] p. 445; [2] Theorem 10.1.5)**.** Given any $(X, \mathcal{S}, p)$ and any σ-subalgebra $\mathcal{A}$ of $\mathcal{S}$, a conditional expectation $\mathscr{E}(\cdot \mid \mathcal{A})$ from $\mathcal{L}^1(X, \mathcal{S}, p)$ to $\mathcal{L}^1(X, \mathcal{A}, p_{\mathcal{A}})$ exists.

Note that uniqueness is not part of the claim in Proposition 2.2 because the conditional expectation is only unique up to measure zero:

Proposition 2.3 ([1] Theorem 16.10 and p. 445; [2] p. 339)**.** If $\mathscr{E}'(\cdot \mid \mathcal{A})$ is another conditional expectation then for any $f \in \mathcal{L}^1(X, \mathcal{S}, p)$ the two $\mathcal{L}^1$-functions $\mathscr{E}(f \mid \mathcal{A})$ and $\mathscr{E}'(f \mid \mathcal{A})$ are equal up to a p-probability zero set.

Different conditional expectations equal up to measure zero are called *versions* of the conditional expectation. It follows that, considered as a map on $L^1(X, \mathcal{S}, p)$, the conditional expectation is unique. We use the notation $\mathbb{E}(\cdot \mid \mathcal{A})$ to denote the conditional expectation $\mathscr{E}(\cdot \mid \mathcal{A})$ when viewed as a map on $L^1(X, \mathcal{S}, p)$. The next proposition states some basic features of the conditional expectations.

Proposition 2.4 ([1] Section 34)**.** A conditional expectation has the following properties:

(i) $\mathbb{E}(\cdot \mid \mathcal{A})$ is a linear map.

(ii) $\mathbb{E}(\cdot \mid \mathcal{A})$ is a projection:

$$\mathbb{E}(\mathbb{E}(f \mid \mathcal{A}) \mid \mathcal{A}) = \mathbb{E}(f \mid \mathcal{A}) \qquad \forall f \in L^1(X, \mathcal{S}, p) \tag{7}$$

(iii) $\mathbb{E}(\cdot \mid \mathcal{A})$ preserves the unit

$$\mathbb{E}(\mathbf{1} \mid \mathcal{A}) = \mathbf{1} \tag{8}$$

(iv) $\mathbb{E}(\cdot \mid \mathcal{A})$ is a $\| \cdot \|_1$-contraction: $\|\mathbb{E}(f \mid \mathcal{A})\|_1 \leq \|f\|_1$ (i.e. $\mathbb{E}(\cdot \mid \mathcal{A})$ is continuous in the $\| \cdot \|_1$-norm topology).

Note that restricted to the Hilbert space $L^2(X, \mathcal{S}, p)$ the conditional expectation $\mathbb{E}(\cdot \mid \mathcal{A})$ is an orthogonal projection on $L^2(X, \mathcal{S}, p)$ with range $L^2(X, \mathcal{A}, p_{\mathcal{A}})$, a closed linear subspace of $L^2(X, \mathcal{S}, p)$.

A deep result of the theory of conditional expectations is that Properties (i)-(iv) in Proposition 2.4 characterize the conditional expectation completely:

Proposition 2.5 ([32], Theorem 3; [11], Corollary 1; [31])**.** Suppose that the map T

$$T\colon \mathcal{L}^1(X,\mathcal{S},p)\to\mathcal{L}^1(X,\mathcal{S},p) \tag{9}$$

is a linear, $\|\cdot\|_1$-contractive projection preserving $\mathbf{1}$. Then there exists a σ-subalgebra $\mathcal{A}$ of $\mathcal{S}$ such that T is a conditional expectation from $\mathcal{L}^1(X,\mathcal{S},p)$ to $\mathcal{L}^1(X,\mathcal{A},p_{\mathcal{A}})$.

3 Conditional probability in terms of conditional expectation

Let $(X,\mathcal{S},p)$ be a probability space, $\mathcal{A}$ be a σ-subalgebra of $\mathcal{S}$. Assume that $\phi'_{\mathcal{A}}$ is a $\|\cdot\|_1$-continuous linear functional on the subspace $L^1(X,\mathcal{A},p_{\mathcal{A}})$ determined by a probability measure $p'_{\mathcal{A}}$ given on the subalgebra $\mathcal{A}$ via integral (cf. equation (2)). What is the extension ϕ' of $\phi'_{\mathcal{A}}$ from $L^1(X,\mathcal{A},p_{\mathcal{A}})$ to a $\|\cdot\|_1$-continuous linear functional on $L^1(X,\mathcal{S},p)$? This question is the general problem of statistical inference, and the answer to it is the concept of conditional probability: One is interested in the expectation values $\phi'(f)$ of random variables f in $L^1(X,\mathcal{S},p)$ that are *not* in $L^1(X,\mathcal{A},p_{\mathcal{A}})$ on condition that the expectation values of functions g that *are* in the narrower set of random variables $L^1(X,\mathcal{A},p_{\mathcal{A}})$ are prescribed (are known) and are given by $\phi'_{\mathcal{A}}(g)$. In general there are many such extensions. *Bayesian* statistical inference yields a particular answer which is based on Bayesian conditioning via the conditional expectation determined by the probability p and the subalgebra $\mathcal{A}$:

Definition 3.1 (Bayesian statistical inference)**.** Let the extension ϕ' of $\phi'_{\mathcal{A}}$ be

$$\phi'(f)\doteq\phi'_{\mathcal{A}}(\mathbb{E}(f\mid\mathcal{A}))\qquad\forall f\in\mathcal{L}^1(X,\mathcal{S},p) \tag{10}$$

where $\mathbb{E}(\cdot\mid\mathcal{A})$ is the $\mathcal{A}$-conditional expectation from $L^1(X,\mathcal{S},p)$ to $L^1(X,\mathcal{A},p_{\mathcal{A}})$.

Note that because $\mathbb{E}(\cdot\mid\mathcal{A})$ is a projection operator on $L^1(X,\mathcal{S},p)$ (Proposition 2.4), ϕ' is indeed an *extension* of $\phi'_{\mathcal{A}}$, and because $\mathbb{E}(\cdot\mid\mathcal{A})$ is $\|\cdot\|_1$-continuous, the extension ϕ' also is $\|\cdot\|_1$-continuous. Thus equation (10) defines an extension ϕ' of $\phi'_{\mathcal{A}}$ indeed.

The notion of conditional *probability* of an event obtains as a special case of Bayesian statistical inference so defined:

Definition 3.2. If $B\in\mathcal{S}$ then its $(\mathcal{A},p'_{\mathcal{A}})$-conditional probability $p'(B)$ is the expectation value $\phi'(\chi_B)$ of its characteristic function χ_B computed using the formula (10) containing the $\mathcal{A}$-conditional expectation:

$$p'(B)\doteq\phi'(\chi_B)=\phi'_{\mathcal{A}}(\mathbb{E}(\chi_B\mid\mathcal{A})) \tag{11}$$

Note that the value $\mathbb{E}(\chi_A\mid\mathcal{A})$ of the conditional expectation $\mathbb{E}(\cdot\mid\mathcal{A})$ on a characteristic function χ_A is *not* a characteristic function: it is only an integrable function. This is why one has to go to function spaces if one wants to define conditional probabilities using conditional expectations the way specified by Definition 3.2.

We now show that both Jeffrey conditionalization and elementary conditionalization via the Bayes rule are particular cases of the conditional probability defined via conditional expectations in the manner given by Definition 3.2. To see this recall first a well-known fact from the theory of conditional expectations:

Proposition 3.3 ([1] p. 446)**.** Let $(X,\mathcal{S},p)$ be a probability space. If the σ-subalgebra $\mathcal{A}$ of $\mathcal{S}$ is generated by a countably infinite partition $\{A_i\}_{i\in\mathbb{N}}$ such that $p(A_i)\neq 0$ $(i=1,\ldots)$, then the conditional expectation (5) can be given explicitly on the characteristic functions of $\mathcal{L}^1(X,\mathcal{S},p)$ as

$$\mathbb{E}(\chi_B \mid \mathcal{A}) = \sum_i \frac{p(B\cap A_i)}{p(A_i)}\chi_{A_i} \qquad \forall B\in\mathcal{S} \tag{12}$$

In particular, if $\mathcal{S}$ has a finite number of elements, then all conditional expectations are of the above form (with a finite summation).

It follows that if state $\phi'_{\mathcal{A}}$ is given on $\mathcal{A}$ by fixing its values $\phi'_{\mathcal{A}}(A_i) = p'_{\mathcal{A}}(A_i)$ on the generating sets A_i, then the $(\mathcal{A},\phi'_{\mathcal{A}})$-conditional probability of $B\in\mathcal{S}$, $B\notin\mathcal{A}$ that Definition 3.2 specifies is

$$\begin{aligned} p'(B) \doteq \phi'(\chi_B) &= \phi'_{\mathcal{A}}(\mathbb{E}(\chi_B \mid \mathcal{A})) && (13)\\ &= \phi'_{\mathcal{A}}\Big(\sum_i \frac{p(B\cap A_i)}{p(A_i)}\chi_{A_i}\Big) && (14)\\ &= \sum_i \frac{p(B\cap A_i)}{p(A_i)}\phi'_{\mathcal{A}}(\chi_{A_i}) && (15)\\ &= \sum_i \frac{p(B\cap A_i)}{p(A_i)}p'_{\mathcal{A}}(A_i) && (16)\end{aligned}$$

which is the Jeffrey conditional rule [23].

Simple Bayesian conditioning is a special case of Jeffrey conditioning: If the Boolean algebra $\mathcal{A}$ is generated by two non-trivial elements $A, A^\perp$ and we take $\phi'_{\mathcal{A}}$ to be the special state on the Boolean algebra $\mathcal{A}$ that takes the values $\phi'_{\mathcal{A}}(A) = 1$ and $\phi'_{\mathcal{A}}(A^\perp) = 0$, then the Jeffrey conditionalization rule (13)-(16) reduces to Bayes' rule:

$$p'(B) = \frac{p(B\cap A)}{p(A)} \tag{17}$$

In light of recovering Bayes' rule this way as a special case of conditioning via conditional expectation it becomes visible that the simple Bayes' rule (17) is slightly deceptive: Bayes' rule gives the impression that it is the probability measure p that gets "upgraded in light of evidence A". But in fact it is the specific probability measure $\phi'_{\mathcal{A}}$ having the particular values $\phi'_{\mathcal{A}}(A) = 1$ on A and $\phi'_{\mathcal{A}}(A^\perp) = 0$ on $A^\perp$ that gets "upgraded" (i.e. extended from the Boolean algebra $\mathcal{A}$ generated by A and $A^\perp$) to a probability measure on the whole σ-algebra $\mathcal{S}$ – the role of the probability measure p is to serve as the background measure with respect to which the upgrading takes place. Thus Bayes' rule conceals somewhat the true logical structure of conditionalization, which is the following:

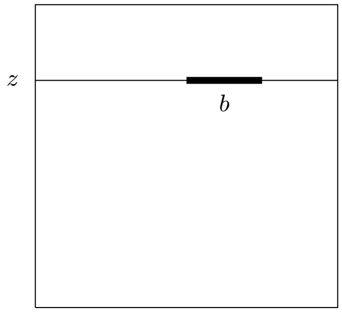
(i) The measure $\phi'_{\mathcal{A}}$ on the subalgebra $\mathcal{A}$ represents the conditioning conditions.

(ii) The extension ϕ' of $\phi'_{\mathcal{A}}$ to the whole algebra $\mathcal{S}$ yields the conditional probability on condition that the values of ϕ' are prescribed on $\mathcal{A}$.

(iii) p is the *fixed* background probability measure with respect to which the conditioned values are obtained via (statistical) inference.

It must be emphasized that conditionalizing using the theory of conditional expectations in the spirit of Definitions 3.1 and 3.2 is much more general than the Jeffrey conditionalization: First, because a general $\mathcal{A}$ is not generated by a countable partition; and in such cases the $\mathcal{A}$-conditional expectation cannot be of the form (12). Second, the $\mathcal{A}$-conditional expectation cannot always be given explicitly, its existence is the corollary of the Radon-Nykodim theorem, which is a non-constructive, pure existence theorem. Third, the formula (12) and hence the Jeffrey conditional, is not defined for events A_i that have zero unconditional probability. But the theory of conditional expectations can handle conditional probabilities with respect to a σ-subalgebra $\mathcal{A}$ that contain some p-probability zero events. The standard example of conditioning with respect to probability zero events is obtaining the conditional probability distribution on the one dimensional slices of the unit square with the Lebesgue measure giving the unconditional probability on the square [1][p. 432]. To illustrate the notion of conditional probability obtained via conditional expectation we describe here this paradigm example briefly.

Example 3.4. Let $(X, \mathcal{S}, p)$ be the probability space with $X = [0,1] \times [0,1]$ the unit square in two dimension, $\mathcal{S}$ the Borel measurable subsets of $[0,1] \times [0,1]$ and $p = l \times l$ the Lebesgue measure on $\mathcal{S}$, where l is the Lebesgue measure on $[0,1]$. Let $C \doteq [0,1] \times \{z\}$ be any horizontal slice of the square at number $z \in [0,1]$ and $B \doteq b \times \{z\}$ be a Borel set of the square with b a Borel set in the slice (see the Fig. 3.4).



What is the conditional probability of B on condition C? Note that slices are measure zero sets in the two dimensional Lebesgue measure and there are an uncountably infinite number of them, hence neither Bayes rule nor Jeffrey conditionalization are applicable. Application of conditionalization via conditional expectation to this situation is possible however and is the following. Consider the σ-algebra $\mathcal{A} \subset \mathcal{S}$ generated by the sets of form $[0,1] \times A$ with A a Borel subset of $[0,1]$. Note that $\mathcal{A}$ contains the slices $[0,1] \times \{z\}$ where z is a number in $[0,1]$; these sets have measure zero in the Lebesgue measure on the square. Then one can check by an elementary calculation that a version of the $\mathcal{A}$-conditional expectation $\mathscr{E}(\cdot \mid \mathcal{A})$ is given explicitly by:

$$\mathscr{E}(f \mid \mathcal{A})(x,y) = \int_0^1 f(x,y)dx \qquad \forall (x,y) \in [0,1] \times [0,1] \tag{18}$$

Inserting the characteristic function χ_B of $B = b \times \{z\}$ in the place of f in eq. (18) one obtains for all $(x,y) \in [0,1] \times [0,1]$:

$$\begin{aligned} \mathscr{E}(\chi_B \mid \mathcal{A})(x,y) &= \int_0^1 \chi_{b\times\{z\}}(x,y)dx && (19) \\ &= \begin{cases} l(b), & \text{if } \ y = z \\ 0, & \text{if } \ y \neq z \end{cases} && (20) \end{aligned}$$

If $p'_{\mathcal{A}}$ is the probability measure on the σ-algebra $\mathcal{A}$ such that

$$\begin{aligned} p'_{\mathcal{A}}(C) &= p'_{\mathcal{A}}([0,1] \times \{z\}) = 1 && (21) \\ p'_{\mathcal{A}}(C^{\perp}) &= p'_{\mathcal{A}}(([0,1] \times \{z\})^{\perp}) = 0 && (22) \end{aligned}$$

then, by definition, the $(\mathcal{A}, p'_{\mathcal{A}})$-conditional probability $p'(b \times \{z\}))$ of B on condition $C = [0,1] \times \{z\}$ (i.e. on condition that $p'_{\mathcal{A}}([0,1] \times \{z\}) = 1$) can be calculated using (19):

$$\begin{aligned} p'(b \times \{z\}) &= p'_{\mathcal{A}}(\mathscr{E}(\chi_{b\times\{z\}} \mid \mathcal{A})) && (23) \\ &= \int_{[0,1]\times[0,1]} \mathscr{E}(\chi_{b\times\{z\}} \mid \mathcal{A}) dp'_{\mathcal{A}} && (24) \\ &= l(b) && (25) \end{aligned}$$

Thus we have obtained that given any one dimensional slice $C = [0,1] \times \{z\}$ at point z across the square, the $(\mathcal{A}, p'_{\mathcal{A}})$-conditional probability of the subset b of that slice on condition that we are on that slice ($p'_{\mathcal{A}}(C) = 1$) is proportional to the length of the subset b. This result is obtained using the technique of conditional expectation with respect to a σ-subalgebra $\mathcal{A}$ some elements of which have probability zero. Again: this result cannot be obtained using Bayes rule or using Jeffrey conditionalization because there is an uncountably infinite number of disjoint slices, hence the algebra $\mathcal{A}$ is not generated by a countable partition.

It should be noted that Jeffrey mentioned the issue of generalization of his rule of conditionalizing in order to include the "continuous case" [23][Section 11.8]. It was also clear to him that "To discuss the matter more rigorously and generally, it is necessary to use the notion of integration over abstract spaces..." [23][p. 177]. But he did not seem to have worked out the general case systematically. Nor did he refer to the theory of conditional expectations, which is precisely the theory developed by Kolmogorov to cover the general case. Expositions of the mathematical theory of conditional expectations, nowadays a standard topic in probability theory, also do not refer to Jeffrey conditionalization. It is not clear to us why the connection has not been made, although, as we have seen, the connection is straightforward.

4 The Bayes accessibility relation in terms of conditional expectations

Definition 4.1. If ϕ is a state in $L^1(X, \mathcal{S}, p)^\sharp$ then we say that ϕ is Bayes accessible for the Bayesian Agent if there exists a σ-subalgebra $\mathcal{A}$ of $\mathcal{S}$ and a state $\psi_{\mathcal{A}}$ in $L^1(X, \mathcal{A}, p_{\mathcal{A}})^\sharp$ such that conditionalizing $\psi_{\mathcal{A}}$ using the conditional expectation

$$\mathbb{E}(\cdot \mid \mathcal{A})\colon L^1(X, \mathcal{S}, p) \to L^1(X, \mathcal{A}, p_{\mathcal{A}}) \tag{26}$$

we obtain ϕ, i.e. if we have

$$\phi(f) = \psi_{\mathcal{A}}(\mathbb{E}(f \mid \mathcal{A})) \qquad \text{for all } f \in L^1(X, \mathcal{S}, p) \tag{27}$$

The Bayesian interpretation of Bayes accessibility of ϕ is straightforward: The probability measure p represents the background knowledge of the Bayesian Agent ($p_{\mathcal{A}}$ is the restriction of p to $\mathcal{A}$). Suppose the Agent wishes to learn the state ϕ. If ϕ is Bayes accessible, then there exists a set of propositions represented by a σ-subalgebra $\mathcal{A}$ of $\mathcal{S}$ such that from the evidence given by the state $\psi_{\mathcal{A}}$ on the subspace $L^1(X, \mathcal{A}, p_{\mathcal{A}})$ determined by the subalgebra $\mathcal{A}$ and by the probability representing the Agent's background measure, the Agent can infer and thus learn the values of ϕ by Bayesian statistical inference, i.e. by Bayesian upgrading using conditional expectations as the conditioning device.

Example 4.2. As an illustration of Bayes accessibility consider the situation described in Example 3.4: Suppose that the Bayesian Agent is told that points have been chosen randomly on the unit square but the Agent is not told what the distribution of the points is, and he assumes that the distribution is given by the uniform (Lebesgue) measure p on the square. (This uniform probability measure represents the Agent's background knowledge.) The Agent is then given the information about what the expectation values of random variables depending on the outcome of choices of the points restricted to Borel sets of horizontal slices $[0,1] \times A$ are (which includes the information about what the distribution of the points is on single horizontal slices). That is to say, the Agent is given, as evidence, a state $\phi'_{\mathcal{A}}$ on the σ-subalgebra $\mathcal{A}$ described in Example 3.4. The Agent is then asked to infer the expectation values $\phi'(f)$ of random variables f on the *whole* square (which includes probabilities $\phi'(\chi_B)$ that the randomly chosen points lie in an *arbitrary* Borel subset B of the square). Following the prescription of Bayesian statistical inference, the Bayesian Agent infers:

$$\phi'(f) = \phi'_{\mathcal{A}}(\mathbb{E}(f \mid \mathcal{A})) \tag{28}$$

where $\mathbb{E}(\cdot \mid \mathcal{A})$ is the $\mathcal{A}$-conditional expectation (cf. equation (18)). In particular, the Agent can infer the conditional probabilities

$$p'(B) = \phi'(\chi_B) = \phi'_{\mathcal{A}}(\mathbb{E}(\chi_B \mid \mathcal{A})) \tag{29}$$

The state ϕ' is thereby Bayes accessible for the Agent because from the evidence represented by the state $\phi'_{\mathcal{A}}$ on the subalgebra $\mathcal{A}$ the Agent can infer ϕ' by Bayesian inference using the theory of conditional expectations determined by his background knowledge. Neither Bayes rule nor Jeffrey conditionalization makes such an inference possible.

To investigate the features of Bayesian learning so defined it is useful to define a general Bayes accessibility relation as follows:

Definition 4.3. If ϕ and ψ are states in $L^1(X, \mathcal{S}, p)^\sharp$ then we say that ϕ is Bayes accessible from ψ (which we denote by $\psi \overset{\mathbb{E}}{\rightsquigarrow} \phi$), if there exists a σ-subalgebra $\mathcal{A}$ of $\mathcal{S}$ such that conditionalizing ψ using the conditional expectation

$$\mathbb{E}(\cdot \mid \mathcal{A})\colon L^1(X, \mathcal{S}, p) \to L^1(X, \mathcal{A}, p_{\mathcal{A}}) \tag{30}$$

we obtain ϕ; i.e. if we have

$$\phi(f) = \psi(\mathbb{E}(f \mid \mathcal{A})) \qquad \text{for all } f \in L^1(X, \mathcal{S}, p) \tag{31}$$

The relation of Definitions 4.1 and 4.3 is straightforward: Since the range of the conditional expectation $\mathbb{E}(\cdot \mid \mathcal{A})$ is the subspace $L^1(X, \mathcal{A}, p_{\mathcal{A}})$ of $L^1(X, \mathcal{S}, p)$, from the perspective of Bayes accessibility of ϕ from ψ only the values of ψ on $L^1(X, \mathcal{A}, p_{\mathcal{A}})$ matter. Thus, if there exists a state $\psi_{\mathcal{A}}$ on the *subspace* $L^1(X, \mathcal{A}, p_{\mathcal{A}})$ from which ϕ can be obtained by conditioning using the conditional expectation $\mathbb{E}(\cdot \mid \mathcal{A})$ (and hence ϕ is Bayes accessible for then Agent), then ϕ can be Bayes accessed from any extension of $\psi_{\mathcal{A}}$ to a state ψ on $L^1(X, \mathcal{S}, p)$. Conversely, if for a state ϕ there exists a state ψ in $L^1(X, \mathcal{S}, p)^\sharp$ such that $\psi \overset{\mathbb{E}}{\rightsquigarrow} \phi$, then ϕ is Bayes accessible for the Bayesian agent in the sense of Definition 4.1 because from the restriction $\psi_{\mathcal{A}}$ of ψ to $L^1(X, \mathcal{A}, p_{\mathcal{A}})$ the Agent can infer ϕ by conditioning. Thus "ϕ is Bayes accessible from *some* ψ" is *equivalent* to "ϕ is Bayes accessible for the Bayesian agent".

The Bayes accessibility (Definition 4.3) defines a two-place relation in the state space $L^1(X, \mathcal{S}, p)^\sharp$, a subset of the dual space $L^1(X, \mathcal{S}, p)^*$ of the space of integrable random variables $L^1(X, \mathcal{S}, p)$. This Bayes accessibility relation is given by the dual $\mathbb{E}(\cdot \mid \mathcal{A})^*$ of conditional expectations $\mathbb{E}(\cdot \mid \mathcal{A})$, where the dual $\mathbb{E}(\cdot \mid \mathcal{A})^*$ of $\mathbb{E}(\cdot \mid \mathcal{A})$ is defined by

$$L^1(X, \mathcal{S}, p)^\sharp \ni \phi \mapsto \mathbb{E}(\cdot \mid \mathcal{A})^*\phi \doteq \phi \circ \mathbb{E}(\cdot \mid \mathcal{A}) \in L^1(X, \mathcal{S}, p)^\sharp \tag{32}$$

Investigating the general properties of general Bayesian learning amounts to determining the features of $\overset{\mathbb{E}}{\rightsquigarrow}$ viewed as a two-place relation in the state space. This is what we do in the present paper.

5 Antisymmetry of the Bayes accessibility relation and failure of strong Bayes connectedness of state spaces

It is obvious that $\overset{\mathbb{E}}{\rightsquigarrow}$ is reflexive: One has to take the identity map on $L^1(X,\mathcal{S},p)$ as the conditional expectation $\mathbb{E}$ to access every state from itself.

Proposition 5.1. The relation $\overset{\mathbb{E}}{\rightsquigarrow}$ is antisymmetric.

Proof. Assume $\psi \overset{\mathbb{E}}{\rightsquigarrow} \phi$ and $\phi \overset{\mathbb{E}}{\rightsquigarrow} \psi$. Then there exist σ-subalgebras $\mathcal{S}_1, \mathcal{S}_2$ of $\mathcal{S}$ and conditional expectations

$$\begin{aligned} \mathbb{E}(\cdot \mid \mathcal{S}_1) \quad &: \quad L^1(X,\mathcal{S},p) \to L^1(X,\mathcal{S}_1,p_1) && (33) \\ \mathbb{E}(\cdot \mid \mathcal{S}_2) \quad &: \quad L^1(X,\mathcal{S},p) \to L^1(X,\mathcal{S}_2,p_2) && (34) \end{aligned}$$

Such that

$$\begin{aligned} \phi(f) &= \psi(\mathbb{E}(f \mid \mathcal{S}_1)) \qquad \forall f \in L^1(X,\mathcal{S},p) && (35) \\ \psi(f) &= \phi(\mathbb{E}(f \mid \mathcal{S}_2)) \qquad \forall f \in L^1(X,\mathcal{S},p) && (36) \end{aligned}$$

Let $\mathbb{E}_1$ and $\mathbb{E}_2$ denote the orthogonal projections on $L^2(X,\mathcal{S},p)$ corresponding to the conditional expectations $\mathbb{E}(\cdot \mid \mathcal{S}_1)$ and $\mathbb{E}(\cdot \mid \mathcal{S}_2)$. Equations (35)-(36) entail then

$$\begin{aligned} \phi(f) &= \psi(\mathbb{E}_1 f) \qquad \forall f \in L^2(X,\mathcal{S},p) && (37) \\ \psi(f) &= \phi(\mathbb{E}_2 f) \qquad \forall f \in L^2(X,\mathcal{S},p) && (38) \end{aligned}$$

Equations (37)-(38) entail

$$\begin{aligned} \phi(f) &= \psi(\mathbb{E}_1\mathbb{E}_2\mathbb{E}_1 f) \qquad \forall f \in L^2(X,\mathcal{S},p) && (39) \\ \psi(f) &= \phi(\mathbb{E}_2\mathbb{E}_1\mathbb{E}_2 f) \qquad \forall f \in L^2(X,\mathcal{S},p) && (40) \end{aligned}$$

and equations (39)-(40) entail that for all $n \in \mathbb{N}$ we have

$$\begin{aligned} \phi(f) &= \psi([\mathbb{E}_1\mathbb{E}_2\mathbb{E}_1]^n f) \qquad \forall f \in L^2(X,\mathcal{S},p) && (41) \\ \psi(f) &= \psi([\mathbb{E}_2\mathbb{E}_1\mathbb{E}_2]^n f) \qquad \forall f \in L^2(X,\mathcal{S},p) && (42) \end{aligned}$$

Since ϕ and ψ are assumed to be $\|\cdot\|_1$-continuous (41)-(42) entail:

$$\begin{aligned} \phi(f) &= \psi(\overset{1}{\lim_{n\to\infty}}([\mathbb{E}_1\mathbb{E}_2\mathbb{E}_1]^n f)) \qquad \forall f \in L^2(X,\mathcal{S},p) && (43) \\ \psi(f) &= \phi(\overset{1}{\lim_{n\to\infty}}([\mathbb{E}_2\mathbb{E}_1\mathbb{E}_2]^n f)) \qquad \forall f \in L^2(X,\mathcal{S},p) && (44) \end{aligned}$$

where $\overset{1}{\lim}$ denotes the limit in the $\|\cdot\|_1$ norm.

Since $\mathbb{E}_1$ and $\mathbb{E}_2$ are projections on the Hilbert space $\mathcal{H} = L^2(X,\mathcal{S},p)$, the limits of the operator sequences $[\mathbb{E}_1\mathbb{E}_2\mathbb{E}_1]^n$ and $[\mathbb{E}_2\mathbb{E}_1\mathbb{E}_2]^n$ exist in the sense of the strong operator topology in the set of all bounded operators $\mathcal{B}(\mathcal{H})$ on $\mathcal{H}$, the limits are the same, and the limit is an element in the lattice $\mathcal{P}(\mathcal{H})$ of all projections on $\mathcal{H}$: it is the greatest lower bound $\mathbb{E}_2 \wedge \mathbb{E}_1$ of the the projections $\mathbb{E}_1$ and $\mathbb{E}_2$ with respect to the standard ordering $\leq$ of projections in $\mathcal{P}(\mathcal{H})$ (Proposition 4.13 in [33]). So for all $f \in L^2(X,\mathcal{S},p)$ we have

$$\begin{aligned} \overset{2}{\lim_{n\to\infty}}[\mathbb{E}_1\mathbb{E}_2\mathbb{E}_1]^n f &= \overset{2}{\lim_{n\to\infty}}[\mathbb{E}_2\mathbb{E}_1\mathbb{E}_2]^n f && (45) \\ &= (\mathbb{E}_2 \wedge \mathbb{E}_1) f && (46) \end{aligned}$$

where $\overset{2}{\lim}$ denotes the limit in the $\|\cdot\|_2$ norm.

Since p is a bounded measure, by Jensen's inequality one has $\|f\|_1 \leq \|f\|_2$; hence the limit of the sequences $[\mathbb{E}_1\mathbb{E}_2\mathbb{E}_1]^n f$ and $[\mathbb{E}_2\mathbb{E}_1\mathbb{E}_2]^n f$ also exists in the $\|\cdot\|_1$ norm, and so for all $f \in L^2(X,\mathcal{S},p)$ we have

$$\begin{aligned} \overset{1}{\lim_{n\to\infty}}[\mathbb{E}_1\mathbb{E}_2\mathbb{E}_1]^n f &= \overset{1}{\lim_{n\to\infty}}[\mathbb{E}_2\mathbb{E}_1\mathbb{E}_2]^n f && (47)\\ &= (\mathbb{E}_2 \wedge \mathbb{E}_1)f && (48)\end{aligned}$$

Equations (43)-(44) together with (47)-(48) entail:

$$\begin{aligned} \phi(f) &= \psi((\mathbb{E}_2 \wedge \mathbb{E}_1)f) \qquad \forall f \in L^2(X,\mathcal{S},p) && (49)\\ \psi(f) &= \phi((\mathbb{E}_1 \wedge \mathbb{E}_2)f) \qquad \forall f \in L^2(X,\mathcal{S},p) && (50)\end{aligned}$$

Since $\mathbb{E}_i \geq (\mathbb{E}_2 \wedge \mathbb{E}_1)$ $(i = 1,2)$, we have

$$\mathbb{E}_i(\mathbb{E}_2 \wedge \mathbb{E}_1) = (\mathbb{E}_2 \wedge \mathbb{E}_1) \qquad i = 1,2 \tag{51}$$

equations (37)-(38) and (49)-(50) entail that for all $f \in L^2(X,\mathcal{S},p)$ we have

$$\begin{aligned} \phi(f) &= \psi((\mathbb{E}_2 \wedge \mathbb{E}_1)f) && (52)\\ &= \phi(\mathbb{E}_2(\mathbb{E}_2 \wedge \mathbb{E}_1)f) && (53)\\ &= \phi((\mathbb{E}_2 \wedge \mathbb{E}_1)f) && (54)\\ &= \psi(f) && (55)\end{aligned}$$

Thus ϕ is equal to ψ on $L^2(X,\mathcal{S},p)$. Since the $L^2(X,\mathcal{S},p)$ is $\|\cdot\|_1$-dense in $L^1(X,\mathcal{S},p)$ and ϕ and ψ are $\|\cdot\|_1$-continuous, ϕ and ψ are equal on $L^1(X,\mathcal{S},p)$. ∎

Antisymmetry of the Bayes accessibility relation $\overset{\mathbb{E}}{\rightsquigarrow}$ entails that state spaces are *not strongly Bayes connected* in general: it is not true that any state ϕ in $L^1(X,\mathcal{S},p)^\sharp$ is Bayes accessible from any other ψ in $L^1(X,\mathcal{S},p)^\sharp$. If strong Bayes connectedness were a feature of a state space then every state could be learned by the Agent by Bayesian upgrading from every other in the same state space: Given any two states ϕ and ψ there would always exist a set of propositions (depending on ϕ of course) such that knowing the values of ψ on elements of that set, the Agent could infer all values of ϕ by conditionalizing ψ (with respect to the fixed background probability measure) on that set of propositions. But antisymmetry of $\overset{\mathbb{E}}{\rightsquigarrow}$ entails that the only probability space that is strongly Bayes connected is the trivial one with $\emptyset$ and X forming $\mathcal{S}$. Thus in non-trivial state spaces Bayesian learning has a certain directedness: if ϕ can be Bayes-learned from ψ, then ψ cannot be Bayes-learned from ϕ. What can be Bayes-learned from some evidence, cannot serve as evidence to Bayes-learn the evidence itself.

6 Are state spaces weakly Bayes connected?

Lack of strong Bayes connectedness of state spaces leads to the following definition:

Definition 6.1. A state space $L^1(X,\mathcal{S},p)^\sharp$ is called *weakly* Bayes connected if for every state ϕ in $L^1(X,\mathcal{S},p)^\sharp$ there exists a state ψ in $L^1(X,\mathcal{S},p)^\sharp$ such that $\psi \neq \phi$ and ϕ is Bayes accessible from ψ.

Are state spaces weakly Bayes connected? There is no general "yes" or "no" answer to this question. We will see that some state spaces are, some others are not weakly Bayes connected. To decide whether a state space is weakly Bayes connected is not a trivial task. For instance we do not know whether a "small enough" state space is weakly Bayes connected (cf. Problem 6.6) (we conjecture that it is not). We show here failure of weak connectedness of state spaces of typical probability theories, and give an example of a weakly Bayes connected state space. To do this, we have to separate the analysis of weak Bayes connectedness into two parts: considering first the state space $L^2(X,\mathcal{S},p)^\sharp$ and then $L^1(X,\mathcal{S},p)^\sharp$. We start with the L^2-theory of Bayes connectedness of state spaces.

Recall that two probability spaces $(X_1,\mathcal{S}_1,p_1)$ and $(X_2,\mathcal{S}_2,p_2)$ are *isomorphic* if there is an invertible map $f: X_1 \to X_2$ such that both f and f^{-1} are measurable, measure preserving maps. A closely related notion is isomorphism modulo zero: $(X_1,\mathcal{S}_1,p_1)$ and $(X_2,\mathcal{S}_2,p_2)$ are *isomorphic modulo* 0 if there exist sets $A_1 \subseteq X_1$ and $A_2 \subseteq X_2$ with $p_1(A_1) = 0 = p_2(A_2)$ such that the probability spaces $(X_1',\mathcal{S}_1,'p_1')$ and $(X_2',\mathcal{S}_2',p_2')$ are isomorphic, where $X_1' = X_1 \setminus A_1$ and $X_2' = X_2 \setminus A_2$, $\mathcal{S}_1'$, and $\mathcal{S}_2'$ are the natural restrictions of the σ-algebras $\mathcal{S}_1$ and $\mathcal{S}_2$ obtained by removing the sets A_1 and A_2, and where p_1' and p_2' are the restrictions of p_1 and p_2 to $\mathcal{S}_1'$, and $\mathcal{S}_2'$. If $(X_1,\mathcal{S}_1,p_1)$ and $(X_2,\mathcal{S}_2,p_2)$ are isomorphic modulo 0, then $L^s(X_1,\mathcal{S}_1,p_1)$ and $L^s(X_2,\mathcal{S}_2,p_2)$ are isometrically isomorphic spaces (because in L^s spaces functions differing on null sets are identified).

Let $(X,\mathcal{S},p)$ be a probability space and $\mathcal{A}$ be a σ-subalgebra of $\mathcal{S}$. We say that $\mathcal{A}$ and $\mathcal{S}$ are equal modulo 0 if $(X,\mathcal{S},p)$ and $(X,\mathcal{A},p)$ are isomorphic modulo 0. $\mathcal{A}$ and $\mathcal{S}$ being not equal modulo zero means that there is a set $A \in \mathcal{S} \setminus \mathcal{A}$ with $p(A) \neq 0$. Note that $L^s(X,\mathcal{A},p)$ is always a closed subspace of $L^s(X,\mathcal{S},p)$ but equality of $\mathcal{A}$ and $\mathcal{S}$ modulo zero implies $L^s(X,\mathcal{S},p) = L^s(X,\mathcal{A},p)$.

The next proposition formulates a condition that is equivalent to the weak Bayes connectedness of L^2 state spaces. Throughout $\mathcal{L}$ denotes the Lebesgue σ-algebra over the reals.

Proposition 6.2. $L^2(X,\mathcal{S},p)^\sharp$ is weakly Bayes connected if and only if there exists no function $f \in L^2(X,\mathcal{S},p)$ which is positive $f > 0$, normalized $\|f\|_2 = 1$, and such that $\mathcal{S}$ and $f^{-1}[\mathcal{L}]$ are equal modulo 0, where $f^{-1}[\mathcal{L}] = \{f^{-1}(A) : A \in \mathcal{L}\}$ with f^{-1} being the inverse image function of f.

Proof. By Riesz's representation theorem ([1][p. 244]) for each state $\phi \in L^2(X,\mathcal{S},p)^\sharp$ there exists a positive, normalized function $f_\phi \in L^2(X,\mathcal{S},p)$ such that

$$\phi(g) = \langle g, f_\phi\rangle \qquad g \in L^2(X,\mathcal{S},p) \tag{56}$$

where $\langle\cdot,\cdot\rangle$ denotes the scalar product in $L^2(X,\mathcal{S},p)$. Conversely: every positive, normalized function f in $L^2(X,\mathcal{S},p)$ defines a state ϕ_f on $L^2(X,\mathcal{S},p)$ by $\phi_f(g) = \langle g, f\rangle$ for all $g \in L^2(X,\mathcal{S},p)$.

Let $\phi,\psi \in L^2(X,\mathcal{S},p)^\sharp$ be two states, f_ϕ and f_ψ be the two functions in $L^2(X,\mathcal{S},p)$ that represent them in the sense of Riesz' representation theorem. If ϕ is Bayes accessible from ψ, then, by definition of Bayes accessibility, there is a σ-subalgebra $\mathcal{A}$ of $\mathcal{S}$ such that

$$\phi(g) = \psi(\mathbb{E}(g \mid \mathcal{A})) \qquad \text{for all } g \in L^2(X,\mathcal{S},p) \tag{57}$$

Denoting by $\mathbb{E}_\mathcal{A}$ the operator on $L^2(X,\mathcal{S},p)$ that represents the conditional expectation $\mathbb{E}(\cdot \mid \mathcal{A})$, and using the Riesz representatives f_ϕ and f_ψ of states ϕ and ψ, equation (57) can be re-written as

$$\langle g, f_\phi\rangle = \langle \mathbb{E}_\mathcal{A} g, f_\psi\rangle \qquad \text{for all } g \in L^2(X,\mathcal{S},p) \tag{58}$$

Since $\mathbb{E}_\mathcal{A}$ is an orthogonal, selfadjoint projection, equation (58) entails

$$\langle g, f_\phi\rangle = \langle \mathbb{E}_\mathcal{A} g, f_\psi\rangle = \langle g, \mathbb{E}_\mathcal{A} f_\psi\rangle \qquad \text{for all } g \in L^2(X,\mathcal{S},p) \tag{59}$$

The equation $\langle g, f_\phi\rangle = \langle g, \mathbb{E}^{\mathcal{A}} f_\psi\rangle$ holds for all $g \in L^2(X, \mathcal{S}, p)$ if and only if $f_\phi = \mathbb{E}^{\mathcal{A}} f_\psi$. Thus we can conclude that if ϕ is Bayes accessible from some state then f_ϕ is in the range of an orthogonal projection $\mathbb{E}^{\mathcal{A}}$ representing a conditional expectation. It follows that if the state ϕ is Bayes accessible from a state *different from* ϕ, then its representing vector f_ϕ must belong to a *proper* closed linear subspace of $L^2(X, \mathcal{S}, p)$ that has the form $L^2(X, \mathcal{A}, p)$. Since the smallest closed linear subspace in $L^2(X, \mathcal{S}, p)$ to which f_ϕ belongs is $L^2(X, f_\phi^{-1}[\mathcal{L}], p)$, state ϕ is Bayes accessible from another state only if $f_\phi^{-1}[\mathcal{L}] \subset \mathcal{S}$ is a proper subalgebra which is not equal to $\mathcal{S}$ modulo 0 (for if $f_\phi^{-1}[\mathcal{L}]$ is a subalgebra equal to $\mathcal{S}$ modulo 0, then $L^2(X, f_\phi^{-1}[\mathcal{L}], p)$ is equal to $L^2(X, \mathcal{S}, p)$). Consequently, $L^2(X, \mathcal{S}, p)^\sharp$ is weakly Bayes connected only if there is no positive and normalized function $f \in L^2(X, \mathcal{S}, p)$ such that $\mathcal{S}$ and $f^{-1}[\mathcal{L}]$ are equal modulo zero.

Conversely, suppose there exists no positive and normalized function f such that $\mathcal{S}$ and $f^{-1}[\mathcal{L}]$ are equal modulo zero. Then for every state ϕ the function f_ϕ that represents ϕ in the sense of the Riesz representation theorem, $L^2(X, f_\phi^{-1}[\mathcal{L}], p)$ is a proper closed linear subspace of $L^2(X, \mathcal{S}, p)$. By Proposition 2.2 there exist then a conditional expectation $\mathbb{E}(\cdot \mid f_\phi^{-1}[\mathcal{L}])$ from $L^2(X, \mathcal{S}, p)$ onto $L^2(X, f_\phi^{-1}[\mathcal{L}], p)$ and (since $L^2(X, f_\phi^{-1}[\mathcal{L}], p)$ is a *proper* subspace) also a positive, normalized function $f' \notin L^2(X, f_\phi^{-1}[\mathcal{L}], p)$ such that $f_\phi = \mathbb{E}(f' \mid f_\phi^{-1}[\mathcal{L}])$. This entails $\langle g, f_\phi\rangle = \langle g, \mathbb{E}(f' \mid f_\phi^{-1}[\mathcal{L}])\rangle$ for all $g \in L^2(X, \mathcal{S}, p)$, which is equivalent to $\phi = \psi \circ \mathbb{E}(\cdot \mid f_\phi^{-1}[\mathcal{L}])$ where ψ is the state in $L^2(X, \mathcal{S}, p)^\sharp$ that is Riesz-represented by function f'. Thus every state in $L^2(X, \mathcal{S}, p)^\sharp$ is obtainable as a conditioned state and so the state space $L^2(X, \mathcal{S}, p)^\sharp$ is weakly Bayes connected. ∎

Lemma 6.3. If there is an injective, positive and normalized function $f \in L^2(X, \mathcal{S}, p)$, then $L^2(X, \mathcal{S}, p)^\sharp$ is *not* weakly Bayes connected.

Proof. If $f : X \to \mathbb{R}$ is injective, then for all $A \in \mathcal{S}$ we have $f^{-1}(f(A)) = A$. This entails $f^{-1}[\mathcal{L}] = \mathcal{S}$ and the statement follows from Proposition 6.2. ∎

To state the next proposition we need to recall the notion of a *standard* probability space. Intuitively, a probability space is *standard* if it is "the sum" of continuous and discrete parts, where the continuous part is (measure theoretically) isomorphic to an interval with the Lebesgue (or Borel) measure on it, and the discrete part is (measure theoretically) isomorphic to a measure space with a σ-algebra that is either finite or is generated by a countably infinite set. To give a more precise definition one has to define the sum (disjoint union) of measure spaces: Let $(X_i, \mathcal{S}_i, p_i)$ for $i < n \in \mathbb{N}$ be finitely many measure spaces and suppose for convenience that the X_i's are disjoint sets. Define a σ-algebra $\mathcal{S}$ on $X = \bigcup_i X_i$ as follows: Take a subset $A \subseteq X$ to be in $\mathcal{S}$ if and only if $A \cap X_i$ belongs to $\mathcal{S}_i$ for all i. Then the map $p : \mathcal{S} \to \mathbb{R}$ defined by

$$p(A) \doteq \sum_i p_i(A \cap X_i) \qquad \text{for all } A \in \mathcal{S} \tag{60}$$

is a measure and the measure space $(X, \mathcal{S}, p)$ is called the disjoint union of the measure spaces $(X_i, \mathcal{S}_i, p_i)$. (For the elementary properties of a disjoint union of measure spaces we refer to [14][section 214K].) A probability space is called standard if it is isomorphic modulo zero to the disjoint union of the Borel or Lebesgue measure spaces of a (possibly empty) interval, and a measure space with a σ-algebra that is either finite or is generated by a countably infinite set (cf. Definition 4.5 in [30]). It is not hard to see that the disjoint union of finitely many standard measure spaces is also standard.

Examples of standard probability spaces include all probability spaces with a finite or countably infinite set of elementary events ("discrete" probability spaces) and the n-dimensional Euclidean spaces

$\mathbb{R}^n$ with probability given by a density function with respect to the Lebesgue measure on $\mathbb{R}^n$. Also included are the probability spaces where X is a compact subset E of $\mathbb{R}^n$ and the probability on E is given by a density with respect to the restriction of the Lebesgue measure to E. These probability spaces cover essentially all applications of probability.

Proposition 6.4. Let $(X, \mathcal{S}, p)$ be a probability space. Then $L^2(X, \mathcal{S}, p)^\sharp$ is *not* weakly Bayes connected in the following (i)-(iii) cases:

(i) $(X, \mathcal{S}, p)$ is generated by a countable set of point masses, i.e. X is finite or countably infinite.

(ii) $(X, \mathcal{S}, p)$ is isomorphic to an interval with the Borel or Lebesgue measure.

(iii) $(X, \mathcal{S}, p)$ is a standard probability space.

Proof. (i) Suppose X is finite or countably infinite. In this case it is clear that there exists a measurable, injective, positive and integrable $f : X \to \mathbb{R}$. By re-normalization we can also assume that f is normalized. Then (i) follows from Lemma 6.3. For later purposes we note that such an f can always be assumed to be bounded and hence to belong to $L^\infty(X, \mathcal{S}, p) \cap L^2(X, \mathcal{S}, p)$. (Take for instance $X = \mathbb{N}$ and $f(n) = \frac{1}{n+1}$.)

(ii) Without loss of generality we can assume that $(X, \mathcal{S}, p)$ is the Lebesgue space $([0, 1], \mathcal{L}, \lambda)$. We wish to apply Lemma 6.3 again. It is easy to see that there is an injective, positive function $f : [0, 1] \to \mathbb{R}$ (take, for instance, the identity function $\mathsf{id}_{[0,1]}$ on $[0, 1]$). Clearly f is measurable and belongs to $L^2([0, 1], \mathcal{L}, \lambda)$. To make it normalized, divide it by $\|f\|_2$. For later purposes we note $\mathsf{id}_{[0,1]} \in L^\infty(X, \mathcal{S}, p) \cap L^2(X, \mathcal{S}, p)$.

(iii) In this case $(X, \mathcal{S}, p)$ is isomorphic modulo zero to a disjoint union of a (possibly empty) interval with Lebesgue or Borel measure and a countable (possibly empty) set of point masses. Take the union of the two injective, positive functions obtained from cases (i) and (ii) and normalize it to length 1. Then the result follows again from Lemma 6.3. ∎

Proposition 6.4 shows that probability spaces are typically not weakly Bayes connected. This leads to the question of whether weakly Bayes connected probability spaces exist at all. We show below that they do by isolating a class of probability spaces which have weakly Bayes connected state spaces. However, the spaces in that class are "very large": Call a probability space $(X, \mathcal{S}, p)$ *significantly large* if it is not isomorphic modulo zero to any space $(X', \mathcal{S}', p')$ with $\mathcal{S}'$ having cardinality less than or equal to the cardinality of the set $\mathcal{L}$ of Lebesgue measurable sets. Significantly large probability spaces exist. Consider for instance the following example. Let X be any uncountable set, and $\mathcal{S}$ be the family of sets $A \subseteq X$ with the property that either A or its complement $X \setminus A$ is countable. Then $\mathcal{S}$ is a σ-algebra of subsets of X, and its cardinality $|\mathcal{S}|$ satisfies $|\mathcal{S}| \geq |X|$. Consider the function $p : \mathcal{S} \to [0, 1]$ defined by $p(A) = 0$ if A is countable and $p(A) = 1$ if A is not countable. Then p is a probability measure on $\mathcal{S}$. If $|X| > 2^{2^{\aleph_0}}$, then $(X, \mathcal{S}, p)$ is significantly large. This is because each p-probability zero set is countable, and removing a countable set does not change the cardinality of X. Recall that $|\mathcal{L}| = 2^{2^{\aleph_0}}$.

The next proposition motivates the definition of significantly large probability spaces.

Proposition 6.5. If $(X, \mathcal{S}, p)$ is significantly large, then $L^2(X, \mathcal{S}, p)^\sharp$ is weakly Bayes connected.

Proof. $\mathcal{S}$ cannot be equal modulo zero to $f^{-1}[\mathcal{L}]$ for any $f \in L^2(X, \mathcal{S}, p)$, because in this case $\mathcal{S}$ would be equal modulo zero to an algebra of cardinality $|f^{-1}[\mathcal{L}]| \leq |\mathcal{L}|$. Thus the result follows directly from Proposition 6.2. ∎

Proposition 6.5 establishes a connection between weak Bayes connectedness of the state space of a probability space and cardinality of the σ-algebra of the propositions over which the Bayesian Agent defines probabilities. This proposition gives a sufficient condition for Bayes connectedness to hold: the Boolean algebra of propositions must be larger than the σ-algebra in the set of (real) numbers with respect to which measurability of the random variables is required. It remains open whether this condition also is necessary however. We conjecture that it is.

From the perspective of Bayesian learning, the sufficient condition for weak Bayes connectedness contained in Proposition 6.5 is very demanding: The Bayesian Agent must be able to comprehend a set of elementary (atomic) propositions cardinality of which is way beyond even that of the continuum. Whether one should allow such an extremely strong concept of Bayesian Agent, is questionable. Proposition 6.5 and its proof also indicate in what way the demanding condition could in principle be weakened: One can read Proposition 6.5 as saying that the cardinality of the σ-algebra in the field in which the random variables take their value and with respect to which measurability of the random variables are demanded give a lower bound on the cardinality of the σ-algebra of random events for which weak Bayes connectedness can hold. To put it differently: the coarser the random variables the smaller the minimal size of the σ-algebra of random events that allows in principle for the corresponding probabilistic theory to be weakly Bayes connected. Thus, as long as one considers real valued random variables in the standard interpretation as real valued maps that are required to be Borel (or Lebesgue) measurable, the state spaces of usual probability theories will *not* be weakly Bayes connected.

Propositions 6.4 and 6.5 also lead to the following open problem.

Problem 6.6. Is there a non-standard probability space $(X, \mathcal{S}, p)$ with cardinality $|\mathcal{S}| = |\mathcal{L}|$ such that its state space $L^2(X, \mathcal{S}, p)^\sharp$ is weakly Bayes connected?

Next, we turn to the question of weak Bayes connectedness of L^1-state spaces.

Proposition 6.7. If $(X, \mathcal{S}, p)$ is a standard probability space, then $L^1(X, \mathcal{S}, p)^\sharp$ is *not* weakly Bayes connected.

Proof. The proof is based on the following idea. Suppose $L^2(X, \mathcal{S}, p)^\sharp$ is not weakly Bayes connected. Then there is a positive, normalized function $f \in L^2(X, \mathcal{S}, p)$ witnessing it: the state ϕ_f is not accessible from any other L^2-state (cf. the proof of Proposition 6.2). Since the dual space of $L^1(X, \mathcal{S}, p)$ is $L^\infty(X, \mathcal{S}, p)$, if f happens to belong to $L^\infty(X, \mathcal{S}, p)$ as well, then f defines a state ϕ in $L^1(X, \mathcal{S}, p)^\sharp$ via

$$\phi(g) = \int fg \, dp \quad \text{for all } g \in L^1(X, \mathcal{S}, p) \tag{61}$$

We claim that such a ϕ is not Bayes accessible from any other state $\psi \in L^1(X, \mathcal{S}, p)^\sharp$; thus this state will witness $L^1(X, \mathcal{S}, p)^\sharp$ not being weakly Bayes connected.

To see that ϕ is not Bayes accessible recall that L^1-states are L^2-states as well because $\|\cdot\|_1 \leq \|\cdot\|_2$ holds due the fact that p is a bounded measure. Consequently $\|\cdot\|_1$-continuity of a state $\psi \in L^1(X, \mathcal{S}, p)^\sharp$ implies $\|\cdot\|_2$-continuity of ψ. Thus, if ϕ were Bayes accessible from $\psi \in L^1(X, \mathcal{S}, p)^\sharp$, then the same ψ (being an $L^2(X, \mathcal{S}, p)$-state as well) would show that the restriction ϕ_f of ϕ to $L^2(X, \mathcal{S}, p)$ is Bayes accessible; a clear contradiction. Thus all one has to prove is that there is a function $f \in L^2(X, \mathcal{S}, p) \cap L^\infty(X, \mathcal{S}, p)$ witnessing that $L^2(X, \mathcal{S}, p)^\sharp$ is not weakly Bayes connected. But this has essentially been done in the proof of Theorem 6.4. ∎

Though probability spaces with finite Boolean algebras are standard hence their state spaces not weakly Bayes connected, we include here another proof of violation of weak Bayes connectedness for the finite case. Wed do this for two reasons: First, because the proof in the finite case shows more explicitly how violation of weak Bayes connectedness occurs. Second, the proof will display an explicit prescription that can be used to obtain a lot of Bayes inaccessible states.

Proposition 6.8. The state space of $L^1(X, \mathcal{S}, p)$ is not weakly Bayes connected if the cardinality of the σ-algebra $\mathcal{S}$ is finite.

Proof. Let $(X_n, \mathcal{S}_n, p_n)$ be a probability space with X_n having $n < \infty$ number of elements and with $\mathcal{S}_n$ being the Boolean algebra of the power set of X_n. Let $L^1(X_n, \mathcal{S}_n, p_n)$ be the associated function space. Without loss of generality we may assume that the probability measure p_n is faithful, i.e. $p_n(\{x_i\}) \neq 0$ for every $i = 1, 2, \dots n$. This is because in the function space $L^1(X_n, \mathcal{S}_n, p_n)$ functions differing on p_n-probability zero sets only are identified, hence if $p_n(\{x_i\}) = 0$ then the characteristic function $\chi_{\{x_i\}}$ of $\{x_i\}$ is the zero element in $L^1(X_n, \mathcal{S}_n, p_n)$. Consequently, $L^1(X_n, \mathcal{S}_n, p_n)$ and $L^1(X_m, \mathcal{S}_m, p_m)$ will be equal, where $(X_m, \mathcal{S}_m, p_m)$ $(m \leq n)$ is obtained from $(X_n, \mathcal{S}_n, p_n)$ by leaving out from X_n the p_n-probability zero events and taking $p_m(\{x_j\}) = p_n(\{x_j\})$ on X_m whenever $p_n(\{x_j\}) \neq 0$. The probability measure p_m is faithful then, and the state space of $L^1(X_n, \mathcal{S}_n, p_n)$ is weakly Bayes connected in the sense of conditional expectations if and only if the state space of $L^1(X_m, \mathcal{S}_m, p_m)$ is. Furthermore, if p_n is faithful, then $L^1(X_n, \mathcal{S}_n, p_n) = \mathcal{L}^1(X_n, \mathcal{S}_n, p_n)$ and $\mathbb{E}(\cdot \mid \mathcal{C}) = \mathscr{E}(\cdot \mid \mathcal{C})$ for any $\mathcal{C}$-conditional expectation. Thus one can carry out the calculations involving conditional expectations $\mathbb{E}(\cdot \mid \mathcal{C})$ in terms of the unique version $\mathscr{E}(\cdot \mid \mathcal{C})$. This will be relied on below.

Since X_n is finite, there exist only a finite number of non-trivial Boolean subalgebras $\mathcal{C}_l$ $(l = 1, 2, \dots, M)$ of $\mathcal{S}_n$; non trivial meaning that $\mathcal{C}_l$ is not $\{\emptyset, X_n\}$ and is not the full Boolean algebra $\mathcal{S}_n$. Each $\mathcal{C}_l$-conditional expectation $\mathscr{E}(\cdot \mid \mathcal{C}_l)$ has the form (cf. Proposition 3.3)

$$\mathscr{E}(\chi_B \mid \mathcal{C}_l) = \sum_k^K \frac{p_n(A_k^l \cap B)}{p_n(A_k^l)} \chi_{A_k^l} \tag{62}$$

where (for any fixed l) A_k^l $(k = 1, 2, \dots K)$ is a partition of $\mathcal{S}_n$ and χ_B is the characteristic function of $B \in \mathcal{S}_n$. Assume that $\psi \overset{\mathbb{E}}{\rightsquigarrow} \phi$. Then for some $\mathcal{C}_l$ we have

$$\phi(\chi_B) = \psi(\mathscr{E}(\chi_B \mid \mathcal{C}_l)) \qquad \text{for all } B \in \mathcal{S}_n \tag{63}$$

Since $\mathcal{C}_l$ is a non-trivial Boolean subalgebra of $\mathcal{S}_n$, at least one A_k^l in $\mathcal{C}_l$ has more than one element from X_n; so if

$$A_k^l = \{x_{k_1}^l, x_{k_2}^l, \dots x_{k_l}^l\} \tag{64}$$

then there exist two distinct elements $x_{k_1}^l, x_{k_2}^l$ in A_k^l. Using (62) and keeping in mind that A_k^l form a partition, we can calculate the probabilities $\phi(\chi_{\{x_{k_1}^l\}})$ and $\phi(\chi_{\{x_{k_2}^l\}})$ as follows:

$$\begin{aligned}
\phi(\chi_{\{x_{k_1}^l\}}) &= \psi(\mathscr{E}(\chi_{\{x_{k_1}^l\}} \mid \mathcal{C}_l)) && (65)\\
&= \psi(\sum_k^K \frac{p_n(A_k^l \cap \{x_{k_1}^l\})}{p_n(A_k^l)} \chi_{A_k^l}) && (66)\\
&= \psi(\frac{p_n(\{x_{k_1}^l\})}{p_n(A_k^l)} \chi_{A_k^l}) && (67)\\
&= \frac{p_n(\{x_{k_1}^l\})}{p_n(A_k^l)} \psi(\chi_{A_k^l}) && (68)
\end{aligned}$$

Clearly, $\phi(\chi_{\{x_{k_2}^l\}})$ can be calculated exactly the same way and we obtain

$$\phi(\chi_{\{x_{k_2}^l\}}) = \frac{p_n(\{x_{k_2}^l\})}{p_n(A_k^l)} \psi(\chi_{A_k^l}) \tag{69}$$

Equations (65)-(68) and (69) entail that if $\psi \overset{\mathbb{E}}{\rightsquigarrow} \phi$ with respect to the conditional expectation $\mathscr{E}(\cdot \mid \mathcal{C}_l)$, then there exist elements $x^l_{k_1} \neq x^l_{k_2}$ such that

$$\phi(\chi_{\{x^l_{k_1}\}}) = p_n(\{x^l_{k_1}\})\frac{\psi(\chi_{A^l_i})}{p_n(A^l_k)} \tag{70}$$

$$\phi(\chi_{\{x^l_{k_2}\}}) = p_n(\{x^l_{k_2}\})\frac{\psi(\chi_{A^l_i})}{p_n(A^l_k)} \tag{71}$$

It follows (recall that p_n is faithful) that if ϕ is such that

$$\frac{\phi(\chi_{\{x_i\}})}{p_n(\{x_i\})} \neq \frac{\phi(\chi_{\{x_j\}})}{p_n(\{x_j\})} \qquad i \neq j;\ 1 \leq i, j \leq n \tag{72}$$

then $\psi \overset{\mathbb{E}}{\rightsquigarrow} \phi$ cannot hold for any of the finite number of conditional expectations $\mathcal{C}_l$.

That for any faithful p_n there exists a ϕ for which (72) holds follows from the following

Lemma 6.9. Let $a_1, a_2, \ldots, a_n$ be real numbers in the semi-closed interval $(0,1]$ such that $\sum_i^n a_i = 1$. Then there exist real numbers $b_1, b_2, \ldots, b_n$ such that

$$b_i \in (0,1] \qquad i = 1,2,\ldots,n \tag{73}$$

$$\sum_i^n b_i = 1 \tag{74}$$

$$\frac{b_i}{a_i} \neq \frac{b_j}{a_j} \quad \text{for all } i \neq j;\ i,j = 1,2,\ldots,n \tag{75}$$

Proof of Lemma: Simple induction: The case $n = 2$ is trivial. Assume (induction hypothesis) that Lemma is true for $n > 2$. Let $a_1, a_2, \ldots, a_{n+1}$ be numbers in $(0,1]$ such that $\sum_i^{n+1} a_i = 1$. Consider the numbers a'_i defined by

$$a'_i \doteq \frac{a_i}{\sum_i^n a_i} \qquad i = 1,2,\ldots,n \tag{76}$$

Then $a'_i \in (0,1]$, and $\sum_i^n a'_i = 1$, so by the induction hypothesis there exist numbers $b_i \in (0,1]$ $(i = 1,2,\ldots,n)$ such that

$$\sum_i^n b_i = 1 \tag{77}$$

$$\frac{b_i}{a'_i} \neq \frac{b_j}{a'_j} \quad \text{for all } i \neq j;\ i,j = 1,2,\ldots,n \tag{78}$$

Which entails

$$\frac{b_i}{a'_i \sum_i^n a_i} \neq \frac{b_j}{a'_j \sum_i^n a_i} \quad \text{for all } i \neq j;\ i,j = 1,2,\ldots,n \tag{79}$$

Hence

$$\frac{b_i}{a_i} \neq \frac{b_j}{a_j} \quad \text{for all } i \neq j;\ i,j = 1,2,\ldots,n \tag{80}$$

Let

$$M = \max_i \left\{ \frac{b_i}{a_i} : i = 1,2,\ldots n \right\} \tag{81}$$

and choose b_{n+1} such that $\frac{b_{n+1}}{a_{n+1}} > M$. Then

$$\frac{b_i}{a_i} \neq \frac{b_j}{a_j} \quad \text{for all } i \neq j;\ i,j = 1,2,\ldots,n+1 \tag{82}$$

Re-normalizing b_i $(i =, 1, 2, \ldots n+1)$ by dividing each b_i by $\sum_i^{n+1} b_i$ in order to satisfy $\sum_i^{n+1} b_i = 1$ preserves (82). So the claim of Lemma is proved. ∎

The proof of Proposition 6.8 also reveals that there exist in fact a large number of probability measures over a finite Boolean algebra that are Bayes inaccessible: There is not only one state ϕ in $L^1(X_n, \mathcal{S}_n, p_n)^\sharp$ which satisfies equation (72) and hence is not Bayes accessible from any state: For all small enough numbers ϵ the states ϕ_ϵ such that

$$|\phi_\epsilon(\chi_{\{x_i\}}) - \phi(\chi_{\{x_i\}})| \leq \epsilon \qquad \text{for all } i = 1, 2, \ldots n \tag{83}$$

also satisfy (72) and thus cannot be obtained via non-trivial conditionalization using conditional expectations from any other state. Thus, we have:

Proposition 6.10. Given a fixed probability measure representing the background degree of belief of the Bayesian Agent on a finite Boolean algebra, there exist an uncountably infinite number of states that are not Bayes accessible for the Bayesian Agent.

A similar proposition can be stated for all *standard* probability spaces, as well: the proof of Propositions 6.4 and 6.7 reveals that the functions $f \in L^\infty(X, \mathcal{S}, p) \cap L^2(X, \mathcal{S}, p)$ witnessing non weak Bayes connectedness of the state spaces $L^s(X, \mathcal{S}, p)^\sharp$ $(s = 1, 2)$ can be chosen infinitely many different ways. This leads to the next proposition.

Proposition 6.11. Given a fixed probability measure representing the background degree of belief of the Bayesian Agent on a standard probability space, there exist an uncountably infinite number of states that are not Bayes accessible for the Bayesian Agent.

We give here an example of a Bayes inaccessible state in the situation described in Example 4.2, which involves a standard probability space:

Example 6.12. Let $(X, \mathcal{S}, p)$ be the space where X is the unit square $(0, 1) \times (0, 1)$, $\mathcal{S}$ is the Lebesgue measurable subsets of X and p is the two-dimensional Lebesgue measure on $\mathcal{S}$. We display a state $\phi \in L^1(X, \mathcal{S}, p)^\sharp$ which cannot be Bayes accessed from any other state.

Real numbers in the open unit interval $(0, 1)$ can be uniquely represented by their decimal expansion with the convention that these must not end with an infinite string of 9's. Let $f : (0, 1) \times (0, 1) \to \mathbb{R}$ be defined by $f(x, y) = z$ if and only if $x = 0.x_1x_2x_3\ldots$, $y = 0.y_1y_2y_3\ldots$ and

$$z = 0.x_1y_1x_2y_2x_3y_3\ldots \tag{84}$$

This f is a measurable one-to-one mapping between the unit interval and the unit square (for the measurability of f we refer to Problem 147 in [26]). It is clear that f is positive, bounded, integrable, so $h \doteq f/\|f\|_1$ is a positive, *normalized*, and injective function. By Lemma 6.3 and the proof of Proposition 6.7 any injective, positive and normalized function $h \in L^2(X, \mathcal{S}, p) \cap L^\infty(X, \mathcal{S}, p)$ gives rise to a state $\phi \in L^1(X, \mathcal{S}, p)^\sharp$ via

$$\phi(g) = \int hg \, dp \quad \text{ for all } g \in L^1(X, \mathcal{S}, p) \tag{85}$$

and Lemma 6.3 and the proof of Proposition 6.7 also show that such a state is not Bayes accessible from any other state.

To sum up: Lack of weak Bayes connectedness of typical state spaces means that there exist probabilities on σ-algebras that are not Bayes accessible for the Bayesian agent in the given framework: Given the agent's background degree of belief on the fixed set of propositions, the agent cannot infer all probability measures via a Bayesian upgrading (using conditional expectations as conditioning device) no matter what evidence he is provided with – if by evidence is meant specifying a probability measure on some proper σ-subalgebra of the fixed set of all propositions. This shows the limits of Bayesian learning under the condition that the evidence available for the Agent is restricted to probability measures on σ-subalgebras of a fixed Boolean σ-algebra. Call this Restricted Evidence Upgrading. Given the limits of Bayesian learning as displayed by Propositions 6.8 6.7 and 6.4 characterizing Restricted Evidence Upgrading, one can ask if the Bayesian Agent can go beyond these limits if the available evidence is not restricted to probability measures on σ-subalgebras of a fixed σ-algebra. This issue will be investigated in section 8. The next section deals with the problem of transitivity of the Bayes accessibility relation.

7 The Bayes accessibility relation is not transitive

Proposition 7.1. The Bayes accessibility relation $\overset{\mathbb{E}}{\rightsquigarrow}$ on $L^1(X,\mathcal{S},p)^\sharp$ is not transitive if the σ-algebra $\mathcal{S}$ has more than 4 elements.

Proof. We show that there exist three states ψ, ϕ and ρ in $L^1(X,\mathcal{S},p)^\sharp$ such that $\psi \overset{\mathbb{E}}{\rightsquigarrow} \phi$ and $\phi \overset{\mathbb{E}}{\rightsquigarrow} \rho$ hold but $\psi \overset{\mathbb{E}}{\not\rightsquigarrow} \rho$. (After this proof, an explicit elementary example of such states will be given, see Example 7.2.)

Let $\mathcal{A}$ and $\mathcal{B}$ be two σ-subalgebras of $\mathcal{S}$ such that there exist elements $A \in \mathcal{A} \setminus \mathcal{B}$ and $B \in \mathcal{B} \setminus \mathcal{A}$ such that $A \cap B = C \neq \emptyset$. Note that if $\mathcal{S}$ has more than four elements, then there exist σ-subalgebras $\mathcal{A}$ and $\mathcal{B}$ of $\mathcal{S}$ with this property: If $\mathcal{S}$ has more than 4 elements, then it has at least 8 elements, and thus there are elements A and B lying in a general position; that is to say, there exist elements A and B for which the following conditions hold:

$$A \not\subseteq B, \quad B \not\subseteq A, \quad A \cap B \neq \emptyset, \quad A \cup B \neq X \tag{86}$$

Let $\mathcal{A}$ and $\mathcal{B}$ be the σ-subalgebras generated by A and B, respectively

$$\mathcal{A} = \{\emptyset, A, A^\perp, X\}, \qquad \mathcal{B} = \{\emptyset, B, B^\perp, X\} \tag{87}$$

Then $\mathcal{A} \neq \mathcal{B}$, and A and B with the assumed property exist.

Let $\mathbb{E}_\mathcal{A}$ and $\mathbb{E}_\mathcal{B}$ be the two projections on the Hilbert space $L^2(X,\mathcal{S},p)$ corresponding to the $\mathcal{A}$-conditional and $\mathcal{B}$-conditional expectations $\mathbb{E}(\cdot \mid \mathcal{A})$ and $\mathbb{E}(\cdot \mid \mathcal{B})$, respectively. The set of all projections on $L^2(X,\mathcal{S},p)$ form an orthocomplemented, orthomodular lattice (see e.g. [25], [33]), where orthomodularity is the property that for any two projections Q and R one has

$$\text{if } Q \leq R \text{ then } R = Q \vee (R \wedge Q^\perp) \tag{88}$$

Applying (88) to $Q = [\mathbb{E}_\mathcal{A} \wedge \mathbb{E}_\mathcal{B}]$ and $R = \mathbb{E}_\mathcal{A}$ and $R = \mathbb{E}_\mathcal{B}$, we obtain

$$\mathbb{E}_\mathcal{A} = [\mathbb{E}_\mathcal{A} \wedge \mathbb{E}_\mathcal{B}] \vee (\mathbb{E}_\mathcal{A} \wedge [\mathbb{E}_\mathcal{A} \wedge \mathbb{E}_\mathcal{B}]^\perp) \tag{89}$$

$$\mathbb{E}_\mathcal{B} = [\mathbb{E}_\mathcal{A} \wedge \mathbb{E}_\mathcal{B}] \vee (\mathbb{E}_\mathcal{B} \wedge [\mathbb{E}_\mathcal{A} \wedge \mathbb{E}_\mathcal{B}]^\perp) \tag{90}$$

Since $[\mathbb{E}_\mathcal{A} \wedge \mathbb{E}_\mathcal{B}]$ is orthogonal to both $(\mathbb{E}_\mathcal{A} \wedge [\mathbb{E}_\mathcal{A} \wedge \mathbb{E}_\mathcal{B}]^\perp)$ and to $(\mathbb{E}_\mathcal{B} \wedge [\mathbb{E}_\mathcal{A} \wedge \mathbb{E}_\mathcal{B}]^\perp)$, and since the join of orthogonal projections is equal to their sum, equations (89)-(90) can be written as

$$\mathbb{E}_\mathcal{A} = [\mathbb{E}_\mathcal{A} \wedge \mathbb{E}_\mathcal{B}] + (\mathbb{E}_\mathcal{A} \wedge [\mathbb{E}_\mathcal{A} \wedge \mathbb{E}_\mathcal{B}]^\perp) \tag{91}$$

$$\mathbb{E}_\mathcal{B} = [\mathbb{E}_\mathcal{A} \wedge \mathbb{E}_\mathcal{B}] + (\mathbb{E}_\mathcal{B} \wedge [\mathbb{E}_\mathcal{A} \wedge \mathbb{E}_\mathcal{B}]^\perp) \tag{92}$$

The product $\mathbb{E}_\mathcal{A}\mathbb{E}_\mathcal{B}$ is a projection if and only if $\mathbb{E}_\mathcal{A}$ and $\mathbb{E}_\mathcal{B}$ commute as operators. Relations (91)-(92) show that $\mathbb{E}_\mathcal{A}$ and $\mathbb{E}_\mathcal{B}$ commute if and only if their parts outside their intersection are orthogonal, i.e. if and only if $(\mathbb{E}_\mathcal{A} \wedge [\mathbb{E}_\mathcal{A} \wedge \mathbb{E}_\mathcal{B}]^\perp)$ and $(\mathbb{E}_\mathcal{B} \wedge [\mathbb{E}_\mathcal{A} \wedge \mathbb{E}_\mathcal{B}]^\perp)$ are orthogonal. But $(\mathbb{E}_\mathcal{A} \wedge [\mathbb{E}_\mathcal{A} \wedge \mathbb{E}_\mathcal{B}]^\perp)$ and

$(\mathbb{E}_{\mathcal{B}} \wedge [\mathbb{E}_{\mathcal{A}} \wedge \mathbb{E}_{\mathcal{B}}]^{\perp})$ are *not* orthogonal because by assumption there exist elements $A \in \mathcal{A}$ and $B \in \mathcal{B}$ such that $A \notin \mathcal{B}$ and $B \notin \mathcal{A}$, so the characteristic functions χ_A and χ_B of the elements $A \in \mathcal{A}$ and $B \in \mathcal{B}$ are in the range of the projections $(\mathbb{E}_{\mathcal{A}} \wedge [\mathbb{E}_{\mathcal{A}} \wedge \mathbb{E}_{\mathcal{B}}]^{\perp})$ and $(\mathbb{E}_{\mathcal{B}} \wedge [\mathbb{E}_{\mathcal{A}} \wedge \mathbb{E}_{\mathcal{B}}]^{\perp})$, respectively, and by the condition $A \cap B = C \neq 0$, for the L^2 scalar product $\langle \chi_A, \chi_B \rangle$ of χ_A and χ_B we have

$$\langle \chi_A, \chi_B \rangle = \int_X \chi_A \chi_B dp = \int_X \chi_C dp = p(C) \neq 0 \tag{93}$$

where we used that p is faithful on $L^2(X, \mathcal{S}, p)$.

Since the product $\mathbb{E}_{\mathcal{A}}\mathbb{E}_{\mathcal{B}}$ is not a projection, it is not equal to any projection $\mathbb{E}_{\mathcal{C}}$ that would represent a conditional expectation $\mathbb{E}(\cdot \mid \mathcal{C})$ defined by a σ-subalgebra $\mathcal{C}$ of $\mathcal{S}$. Thus for any such projection $\mathbb{E}_{\mathcal{C}}$ there is an element $f \in L^2(X, \mathcal{S}, p) \subset L^1(X, \mathcal{S}, p)$ such that

$$\mathbb{E}_{\mathcal{A}}\mathbb{E}_{\mathcal{B}} f \neq \mathbb{E}_{\mathcal{C}} f \tag{94}$$

The state space $L^1(X, \mathcal{S}, p)^\sharp$ is separating: for any $f \neq g$ in $L^1(X, \mathcal{S}, p)$, there exists a state ψ in $L^1(X, \mathcal{S}, p)^\sharp$ such that $\psi(f) \neq \psi(g)$, so there is a state ψ such that

$$\psi(\mathbb{E}_{\mathcal{A}}\mathbb{E}_{\mathcal{B}} f) \neq \psi(\mathbb{E}_{\mathcal{C}} f) \tag{95}$$

It follows that defining states ϕ and ρ by

$$\begin{aligned} \phi(f) &\doteq \psi(\mathbb{E}(f \mid \mathcal{A})) &(96)\\ \rho(f) &\doteq \phi(\mathbb{E}(f \mid \mathcal{B})) &(97) \end{aligned}$$

we have $\psi \overset{\mathbb{E}}{\rightsquigarrow} \phi$ and $\phi \overset{\mathbb{E}}{\rightsquigarrow} \rho$ but $\psi \overset{\mathbb{E}}{\rightsquigarrow} \rho$ does not hold. ∎

We illustrate failure of transitivity of the Bayes accessibility relation with the following example.

Example 7.2. We give an explicit example of a probability space and three states ϕ, ψ and ρ in its state space such that $\phi \overset{\mathbb{E}}{\rightsquigarrow} \psi$, $\psi \overset{\mathbb{E}}{\rightsquigarrow} \rho$ but $\phi \overset{\mathbb{E}}{\rightsquigarrow} \rho$ does not hold.

Let $X_3 = \{x_1, x_2, x_3\}$, $\mathcal{S}_3$ be the power set of X_3, and p_3 be the uniform measure on X_3: $p_3(\{x_i\}) = \frac{1}{3}$ $(i = 1, 2, 3)$. There are three non-trivial Boolean subalgebras of $\mathcal{S}$, they are:

$$\begin{aligned} \mathcal{C}_1 &= \{\emptyset, \{x_1\}, \{x_2, x_3\}, X_3\} &(98)\\ \mathcal{C}_2 &= \{\emptyset, \{x_2\}, \{x_1, x_3\}, X_3\} &(99)\\ \mathcal{C}_3 &= \{\emptyset, \{x_3\}, \{x_1, x_2\}, X_3\} &(100) \end{aligned}$$

$\mathbb{E}(\cdot \mid \mathcal{C}_1)$, $\mathbb{E}(\cdot \mid \mathcal{C}_2)$ and $\mathbb{E}(\cdot \mid \mathcal{C}_3)$ are the three conditional expectations from $L^1(X_3, \mathcal{S}_3, p_3)$ to $L^1(X_3, \mathcal{C}_i, p_3)$ $(i = 1, 2, 3)$. These conditional expectations are given on the characteristic functions χ_B of $B \in \mathcal{S}$ by

$$\begin{aligned} \mathbb{E}(\chi_B \mid \mathcal{C}_1) &= \frac{p(\{x_1\} \cap B)}{p(\{x_1\})}\chi_{\{x_1\}} + \frac{p(\{x_2, x_3\} \cap B)}{p(\{x_2, x_3\})}\chi_{\{x_2,x_3\}} &(101)\\ \mathbb{E}(\chi_B \mid \mathcal{C}_2) &= \frac{p(\{x_2\} \cap B)}{p(\{x_2\})}\chi_{\{x_2\}} + \frac{p(\{x_1, x_3\} \cap B)}{p(\{x_1, x_3\})}\chi_{\{x_1,x_3\}} &(102)\\ \mathbb{E}(\chi_B \mid \mathcal{C}_3) &= \frac{p(\{x_3\} \cap B)}{p(\{x_3\})}\chi_{\{x_3\}} + \frac{p(\{x_1, x_2\} \cap B)}{p(\{x_1, x_2\})}\chi_{\{x_1,x_2\}} &(103) \end{aligned}$$

Let ϕ be the state on $L^1(X_3, \mathcal{S}_3, p_3)$ defined by the following probabilities:

$$\phi(\chi_{\{x_1\}}) \doteq \frac{1}{2} \qquad \phi(\chi_{\{x_2\}}) \doteq \frac{1}{6} \qquad \phi(\chi_{\{x_3\}}) \doteq \frac{2}{6} \tag{104}$$

Let ψ and ρ be the states on $L^1(X, \mathcal{S}, p)$ which are *defined by*

$$\psi(f) \doteq \phi(\mathbb{E}(f \mid \mathcal{C}_1) \tag{105}$$

$$\rho(f) \doteq \psi(\mathbb{E}(f \mid \mathcal{C}_2) \tag{106}$$

So $\phi \overset{\mathbb{E}}{\rightsquigarrow} \psi$ and $\psi \overset{\mathbb{E}}{\rightsquigarrow} \rho$ hold by the very definition of these states. We claim that $\phi \overset{\mathbb{E}}{\rightsquigarrow} \rho$ does not hold however. To see this, one can explicitly compute the values of ρ, they are:

$$\rho(\chi_{\{x_1\}}) = \frac{3}{8} \qquad \rho(\chi_{\{x_2\}}) = \frac{1}{4} \qquad \rho(\chi_{\{x_3\}}) = \frac{3}{8} \tag{107}$$

One also can compute explicitly the values of $\phi(\mathbb{E}(\chi_B \mid \mathcal{C}_i))$, for $B \in L^1(X_3, \mathcal{S}_3, p_3)$ for each $\mathcal{C}_i$-conditional expectation $i = 1, 2, 3$: For the elementary event $B = \{x_1\}$ these values are:

$$\phi(\mathbb{E}(\chi_{\{x_1\}} \mid \mathcal{C}_1)) = \frac{1}{2} \tag{108}$$

$$\phi(\mathbb{E}(\chi_{\{x_1\}} \mid \mathcal{C}_2)) = \frac{5}{12} \tag{109}$$

$$\phi(\mathbb{E}(\chi_{\{x_1\}} \mid \mathcal{C}_3)) = \frac{1}{3} \tag{110}$$

Thus

$$\phi(\mathbb{E}(\chi_{\{x_1\}} \mid \mathcal{C}_i)) \neq \rho(\chi_{\{x_1\}}) \qquad \text{for all } i = 1, 2, 3 \tag{111}$$

and since $\mathcal{C}_i$, $(i = 1, 2, 3)$ are the only non-trivial Boolean subalgebras of $\mathcal{S}_3$, on can conclude that $\phi \overset{\mathbb{E}}{\rightsquigarrow} \rho$ does not hold.

Failure of transitivity of $\overset{\mathbb{E}}{\rightsquigarrow}$ in means that “There is no Bayesian royal road to learning” in general: Even if a state ρ can be learned from another state ϕ by several successive Bayesian upgradings using conditional expectations, this step-by-step learning cannot be shortcut in general by a single Bayesian upgrading on a single evidence.

The proof of failure of transitivity shows that lack of transitivity of the Bayes accessibility relation is due to the fact that the Hilbert space projections representing the conditional expectations on the square integrable random variables do not commute as operators. This non-commutativity has been noticed by a number of authors in connection with upgrading using Jeffrey conditionalization [13], [7], [36], [37], [10], [28], [39], and it has been subject of analysis of a string of recent papers [41], [15], [40], [42]. Now we see that it is a general feature of conditionalizing via conditional expectations.

Analyzing the source of non-commutativity of Jeffrey conditionalization, Weisberg finds it in what he calls the “rigidity” of upgrading:

> “Strict Conditionalization and Jeffrey Conditionalization are both rigid, meaning that they preserve the conditional probabilities on the evidence. When we apply Strict Conditionalization to evidence E, $q(H|E) = p(H|E)$. Similarly, if we apply Jeffrey Conditionalization to the partition $\{E_i\}$, then $q(H|E_i) = p(H|E_i)$ for each E_i.” [41][p. 806]

(cf. also [42][p. 125]). Formulated in the terminology of the present paper, rigidity is simply the feature of upgrading that the state ϕ and the conditioned state $\phi \circ \mathbb{E}(\circ \mid \mathcal{A})$ coincide when restricted to the subspace $L^1(X, \mathcal{A}, p_{\mathcal{A}})$. In other words, rigidity is the feature that the Bayesian learning consists in *extending* the probability measure that represents the evidence. Weisberg’s claim that this requirement is indeed crucial and responsible for non-commutativity can be strengthened by extending it to the general situation by the following reasoning:

Suppose one wishes to preserve commutativity of learning by replacing upgrading via conditional expectation with an upgrading procedure that is characterized by a map T on the set of integrable functions $L^1(X, \mathcal{S}, p)$ which is *not* assumed to be a conditional expectation: Given a state ϕ, the composition $\phi \circ T$ would be the upgraded ("conditionalized") state under the new upgrading rule. It is a minimal requirement that (i) T is linear and $\| \cdot \|_1$-continuous (otherwise $\phi \circ T$ is not a state); also (ii) T should be unit preserving $T\mathbf{1} = \mathbf{1}$ (i.e. preserve the characteristic function χ_X of the whole set X of elementary random events) otherwise $\phi \circ T$ would not be normalized hence again not a state. The crucial observation is that if in addition to (i) and (ii) the upgrading rule is rigid, i.e. it has the feature that there is a (closed linear) subspace H in $L^1(X, \mathcal{S}, p)$ such that the T-conditioned states $\phi \circ T$ coincide with the pre-conditioned states ϕ on H for every ϕ, then T must be the identity on H. But then by the deep result characterizing conditional expectations (Proposition 2.5) the subspace H must be of the form $L^1(X, \mathcal{A}, p_{\mathcal{A}})$ with some σ-subalgebra $\mathcal{A}$ of $\mathcal{S}$ and T must be the conditional expectation from $L^1(X, \mathcal{S}, p)$ onto $L^1(X, \mathcal{A}, p_{\mathcal{A}})$. In short: the rigidity requirement on upgrading forces the upgrading to be a conditional expectation quite generally. This entails that the representatives of rigid upgradings will be orthogonal projections on the Hilbert space $L^2(X, \mathcal{S}, p)$ determined by the given probability theory, and these projections do not commute in general. This in turn entails failure of transitivity of the Bayes accessibility relation by Proposition 7.1.

Failure of transitivity of upgrading (equivalently: non-commutativity of upgrading) is generally regarded as intuitively problematic because, as the standard reasoning goes, the result of upgrading a probability measure on the basis of some evidence should not depend on the order in which elements of the evidence is presented to the Agent. In Weisberg's formulation:

> "[A] commonly held desideratum is *commutativity*, the view that the order in which information is learned should not matter to the conclusions we ultimately draw, provided the same total information is collected." [41][p. 794]

Whether the desideratum, which we shall refer to as the "Norm of Epistemic Commutativity", is reasonable as a general requirement, can be debated. We do not wish to argue for or against it here. We will argue however that the failure of commutativity of upgrading in general Bayesian learning which Proposition 7.1 is based on should not be interpreted as a violation of the Norm of Epistemic Commutativity. Quite on the contrary: We will show that, identifying "information" with evidence in Bayesian learning, a careful articulation of the Norm of Epistemic Commutativity in connection with Bayesian learning shows that the Bayesian statistical inference based on conditional expectations satisfies this norm.

Recall that the evidence in Bayesian learning is a single state (probability measure) regarded as defined on a subspace $L^1(X, \mathcal{A}, p)$ of $L^1(X, \mathcal{S}, p)$, where $\mathcal{A}$ is a σ-subalgebra of the σ-algebra $\mathcal{S}$ that represents the whole set of propositions (section 5). As was seen in the proof of Proposition 7.1, the non-commutativity of Bayesian learning is the following phenomenon: Let $\mathcal{A}$ and $\mathcal{B}$ be two Boolean σ-subalgebras of $\mathcal{S}$. The Agent infers state $\psi \circ \mathbb{E}(\cdot \mid \mathcal{A})$ from state ψ (viewed as evidence on $L^1(X, \mathcal{A}, p)$) by upgrading ψ via the conditional expectation $\mathbb{E}(\cdot \mid \mathcal{A})$. Then the Agent considers the *restriction* of the upgraded state $\psi \circ \mathbb{E}(\cdot \mid \mathcal{A})$ to the subspace $\mathbb{E}(\cdot \mid \mathcal{B})$ as new evidence, and upgrades this state, this time by the conditional expectation $\mathbb{E}(\cdot \mid \mathcal{B})$, to infer state $\psi \circ \mathbb{E}(\cdot \mid \mathcal{B}) \circ \mathbb{E}(\cdot \mid \mathcal{A})$. If the Agent does these upgradings in the reversed order, obtaining state $\psi \circ \mathbb{E}(\cdot \mid \mathcal{A}) \circ \mathbb{E}(\cdot \mid \mathcal{B})$, then in general (i.e. for some states ψ) we have

$$\psi \circ \mathbb{E}(\cdot \mid \mathcal{B}) \circ \mathbb{E}(\cdot \mid \mathcal{A}) \neq \psi \circ \mathbb{E}(\cdot \mid \mathcal{A}) \circ \mathbb{E}(\cdot \mid \mathcal{B}) \tag{112}$$

Should this be viewed as violation of the Norm of Epistemic Commutativity? We argue that it should not. Our line of reasoning is the following: We take the uncontroversial inference rule of elementary propositional logic (modus ponens), and we will repeat with this inference rule the (analogue of the) above steps of Bayesian statistical inference that has led to non-commutativity. It will be seen that modus ponens also violates commutativity exactly in the sense in which Bayesian statistical inference does. Then we articulate the Norm of Epistemic Commutativity in terms of elementary classical logic using modus ponens, and show that the modus ponens *does* satisfy the Norm. Finally, we formulate the Norm of Epistemic Commutativity for the Bayesian statistical inference along the lines of the formulation for modus ponens, and refer to some theorems on conditional expectations which show that the Norm also holds for Bayesian statistical inference.

Consider the usual inference rule (modus ponens) in elementary propositional logic. In the terminology of the Boolean algebra $\mathcal{S}$ determined by a zeroth order language of a propositional logic, if $\mathcal{A}$ is a set of propositions from $\mathcal{S}$ that does not contain a contradiction, then the set of all propositions one can infer from $\mathcal{A}$ using modus ponens is the proper filter $F(\mathcal{A})$ in $\mathcal{S}$ generated by $\mathcal{A}$ [18]. Let $\mathcal{A}$ and $\mathcal{B}$ be two sets of propositions in $\mathcal{S}$. Regarding $\mathcal{A}$ as evidence, one can infer $F(\mathcal{A})$ from this information. Following the logic of Bayesian statistical inference, one can then restrict the inferred information $F(\mathcal{A})$ to the part that is contained in the set $\mathcal{B}$. This way one obtains a new evidence set $F(\mathcal{A}) \cap \mathcal{B}$. This new evidence can then be used to infer $F(F(\mathcal{A}) \cap \mathcal{B})$ in a second inference move. Performing two such inferences in the reversed order, starting with $\mathcal{B}$, one obtains $F(F(\mathcal{B}) \cap \mathcal{A})$. It is clear that in general one has:

$$F(F(\mathcal{A}) \cap \mathcal{B}) \neq F(F(\mathcal{B}) \cap \mathcal{A}) \tag{113}$$

This is precisely the kind of non-commutativity displayed by the non-commutativity of the upgrading via conditionalizing using conditional expectation (expressed by inequality (112)). But the inequality (113) does not represent a violation of the Norm of Epistemic Commutativity by the inference in classical propositional logic because the Norm of Epistemic Commutativity is not expressed by the equality $F(F(\mathcal{A})\cap\mathcal{B}) = F(F(\mathcal{B})\cap\mathcal{A})$; rather, compliance of modus ponens with the Norm of Epistemic Commutativity is expressed by the following proposition:

Proposition 7.3. Let $\mathcal{A}$ be any set of propositions and let $\mathcal{C}_i$ $(i = 1, \ldots)$ be any sequence of subsets of $\mathcal{A}$ such that $\mathcal{C}_i \subset \mathcal{C}_j$ if $i < j$ and $\cup_i \mathcal{C}_i = \mathcal{A}$. Then we have:

$$F(\cup_i \mathcal{C}_i) = \cup_i F(\mathcal{C}_i) \tag{114}$$

This proposition says that given any set $\mathcal{A}$ of total information, we can take *any* part $\mathcal{C}_1$ of that total information; can draw the consequences of the information contained in this partial set $\mathcal{C}_1$, and we can then add to this partial information further information from $\mathcal{A}$ step by step *in any order*, each time drawing the consequences of the expanded but still partial set $\mathcal{C}_i$. Then the sum of the consequences drawn from the partial information sets is always the same and is equal to the consequences we can draw from the total information.

Proof.[of Proposition 7.3]

The containment $F(\cup_i \mathcal{C}_i) \supseteq \cup_i F(\mathcal{C}_i)$ is obvious. If $A \in F(\cup_i \mathcal{C}_i)$ then there is a *finite* set $\{A_1, A_2, \ldots A_n\} \subset \cup_i \mathcal{C}_i$ such that $A_1 \cap A_2 \cap \ldots \cap A_n \subseteq A$ (the set $\{A_1, A_2, \ldots A_n\}$ contains the premises in $\mathcal{A}$ from which A can be deduced). Since $\{A_1, A_2, \ldots A_n\}$ is finite and $\cup_i \mathcal{C}_i = \mathcal{A}$, there is a j such that $\{A_1, A_2, \ldots A_n\} \subseteq \mathcal{C}_j$. But then $A \in F(\mathcal{C}_j)$ and thus also $A \in \cup_i F(\mathcal{C}_j)$. ∎

For Bayesian statistical inference the analogue of the epistemic commutativity of the modus ponens expressed by Proposition 7.3 holds: Consider the sequence $\mathcal{C}_i$ $(i = 1, \ldots)$ of σ-subalgebras of $\mathcal{S}$ such that $\mathcal{C}_i \subset \mathcal{C}_j$ for $i < j$, and let $\mathcal{A}$ be the σ-algebra generated by $\cup_i \mathcal{C}_i$ (denoted by $\mathcal{C}_i \uparrow \mathcal{A}$). If $\psi \in L^1(X, \mathcal{S}, p)$ is a state that is viewed as evidence on $L^1(X, \mathcal{A}, p)$ and one performs a Bayesian upgrading $\psi \circ \mathbb{E}(\cdot \mid \mathcal{A})$, then one could perform this learning from the "total information" also step-by-step by doing the upgradings $\psi \circ \mathbb{E}(\cdot \mid \mathcal{C}_i)$ for all $i = 1, 2, \ldots$. This is because $\mathbb{E}(f \mid \mathcal{C}_i)$ converges to $\mathbb{E}(f \mid \mathcal{A})$ in $\| \cdot \|_1$-norm (for all f in $L^1(X, \mathcal{S}, p)$) as $\mathcal{C}_i \uparrow \mathcal{A}$ (Theorem 35.6 in [1]). Note that in this series of upgradings any two conditional expectations also commute as operators because $\mathcal{C}_i \subset \mathcal{C}_j$ (for $i < j$) entails $\mathbb{E}(\mathbb{E}(\cdot \mid \mathcal{C}_j) \mid \mathcal{C}_i) = \mathbb{E}(\cdot \mid \mathcal{C}_i)$ ("tower property" of conditional expectations [1][Theorem 34.4])), which in turn entails the commutativity of the Hilbert space projections representing $\mathbb{E}(\cdot \mid \mathcal{C}_j)$ and $\mathbb{E}(\cdot \mid \mathcal{C}_i)$.

To summarize: Non-commutativity of the conditional expectations determined by different conditioning Boolean subalgebras does not indicate failure of the Norm of Epistemic Commutativity. One can in fact show that the Norm of Epistemic Commutativity holds for Bayesian upgrading based on the technique of conditional expectations. Thus while one might have very good reasons to look for upgradings different from the general Bayesian one (which includes Jeffrey conditionalization), violation of the Norm of Epistemic Commutativity does not seem to be one of the justified reasons.

8 Bayes connectability in terms of conditional expectations

Failure of weak Bayes connectedness of state spaces displays the limits of Bayesian learning under Limited Evidence Upgrading: the evidence available for the Agent is limited to probability measures on σ-subalgebras of the fixed σ-algebra on which the Agent's background probability is given. It is natural however to ask what the Agent can learn via conditionalization using conditional expectations if he is allowed access to potentially unlimited evidence. To investigate this question, we define first the concept of extensions of state spaces.

Definition 8.1. We say that the state space $L^1(X', \mathcal{S}', p')^\sharp$ extends the state space $L^1(X, \mathcal{S}, p)^\sharp$ if the following hold:

(i) There is a measurable, measure preserving map $h : X' \to X$ such that its inverse image function f^{-1} induces a σ-algebra embedding $h^{-1} : \mathcal{S} \to \mathcal{S}'$.

(ii) The σ-algebra embedding preserves the probability: For all $A \in \mathcal{S}$ we have $p(A) = p'(h^{-1}(A))$.

If (i)-(ii) hold, then the embedding of $\mathcal{S}$ into $\mathcal{S}'$ via f^{-1} can be lifted to an isometric embedding $\bar{h} : L^1(X, \mathcal{S}, p) \to L^1(X', \mathcal{S}', p')$ by defining $\bar{h}$ in the natural way: For a function $f \in L^1(X, \mathcal{S}, p)$ let $\bar{h}(f) = f \circ h$ (see the figure below). Since h is measurable, we have $\bar{h}(f) \in L^1(X', \mathcal{S}', p')$.

$$\begin{tikzcd}
X \arrow[r, "f"] & \mathbb{R} \arrow[d, leftrightarrow, "id"] \\
X' \arrow[u, "h"] \arrow[r, dashed, "f \circ h"'] & \mathbb{R}
\end{tikzcd}$$

Note that $\bar{h}$ is isometric because

$$\|\bar{h}(f)\|_1 = \int_{X'} |f \circ h| dp' = \int_X |f| dp = \|f\|_1 \tag{115}$$

The image of $\bar{h}$ is thus a closed subspace in $L^1(X', \mathcal{S}', p')$; hence for each state $\phi \in L^1(X, \mathcal{S}, p)^\sharp$ there is a corresponding state $\bar{\phi} \in \bar{h}\big(L^1(X, \mathcal{S}, p)\big)^\sharp$ such that

$$\bar{\phi}\big(\bar{h}(f)\big) = \phi(f) \quad \text{for all } f \in L^1(X, \mathcal{S}, p) \tag{116}$$

By the Hahn–Banach theorem $\bar{\phi}$ extends to a continuous linear functional $\phi' \in L^1(X', \mathcal{S}', p')^\sharp$. Notice that $\bar{\phi}$ can have many such extensions, in general. Any such ϕ' is called an extension of ϕ.

Definition 8.2. The state space $L^1(X, \mathcal{S}, p)^\sharp$ is called weakly Bayes *connectable* if there is a state space extension $L^1(X', \mathcal{S}', p')^\sharp$ such that each $\phi \in L^1(X, \mathcal{S}, p)^\sharp$ has an extension $\phi' \in L^1(X', \mathcal{S}', p')^\sharp$ which is Bayes accessible from *some* $\psi \in L^1(X', \mathcal{S}', p')^\sharp$, $\psi \neq \phi'$.

The above definition is a significantly generalized version of the definition given by Diaconis and Zabell [6][section 2.1]. Accordingly, the proposition below generalizes Theorem 2.1 in [6].

Proposition 8.3. State spaces $L^1(X, \mathcal{S}, p)^\sharp$ are weakly Bayes connectable.

Proof. Let ϕ be a state in $L^1(X, \mathcal{S}, p)^\sharp$. We have to construct a state space extension $L^1(X', \mathcal{S}', p')^\sharp$ such that the extension $\phi' \in L^1(X', \mathcal{S}', p')^\sharp$ of $\phi \in L^1(X, \mathcal{S}, p)^\sharp$ is Bayes accessible from some $\psi \in L^1(X', \mathcal{S}', p')^\sharp$. The idea of the proof is the following. We take as the extension of $L^1(X, \mathcal{S}, p)$ the product of $L^1(X, \mathcal{S}, p)$ with another probability space $(Y, \mathcal{B}, q)$. The product structure defines a conditional expectation to the components in the product in a canonical manner, and it also makes possible to extend states defined on the components in different ways. We display two extensions of ϕ that will be shown to be related to each other via conditioning with respect to the canonical conditional expectation.

Let $(Y, \mathcal{B}, q)$ be the Lebesgue measure space over the unit interval and consider the usual product space

$$(X \times Y, \mathcal{S} \otimes \mathcal{B}, p \times q) \tag{117}$$

where $p \times q$ is the product measure: $(p \times q)(A \times B) = p(A)q(B)$.

The function

$$X \times Y \ni (x, y) \mapsto h(x, y) \doteq x \in X \tag{118}$$

is a measurable, measure preserving map and its inverse image induces a σ-algebra embedding $h^{-1} : \mathcal{S} \to \mathcal{S}'$, since for all $A \in \mathcal{S}$ we have

$$h^{-1}(A) \quad = \quad A \times Y \in \mathcal{S}' \tag{119}$$

$$p(A) \quad = \quad p'(A \times Y) = p(A)q(Y) \tag{120}$$

h can be lifted to an isometric embedding $\bar{h} : L^1(X, \mathcal{S}, p) \to L^1(X', \mathcal{S}', p')$ by the definition

$$\bar{h}(f) = \bar{f} = f \circ h \qquad f \in L^1(X, \mathcal{S}, p) \tag{121}$$

In what follows, for notational convenience we write $L^1(X)$ and $L^1(X \times Y)$ instead of the longer $L^1(X, \mathcal{S}, p)$ and $L^1(X', \mathcal{S}', p')$, and, to make notation easier to read, we write $\int dx$ and $\int dy$ instead of $\int dp$ and $\int dq$.

The general definition of extension of state spaces (Definition 8.1) in the present context means that if $\phi \in L^1(X)^\sharp$ is a state, then $\phi' \in L^1(X \times Y)^\sharp$ is its extension if for all $f \in L^1(X)$ we have

$$\phi'(\bar{f}) = \phi(f) \tag{122}$$

If $\alpha \in L^1(Y)^\sharp$ is a state in the second component of the product space (117), then we define the α-extension of ϕ (denoted by ϕ_α) to be a state in $L^1(X \times Y)^\sharp$ by setting for any $f \in L^1(X \times Y)$

$$\phi_\alpha(f) \doteq \alpha\big(y \mapsto \phi(x \mapsto f(x, y))\big) \tag{123}$$

Then ϕ_α is an extension of ϕ because for each $f \in L^1(X)$ we have

$$\begin{aligned} \phi_\alpha(\bar{f}) &= \phi_\alpha(f \circ h) = \alpha\big(y \mapsto \phi(x \mapsto (f \circ h)(x,y))\big) && (124) \\ &= \alpha(y \mapsto \phi(x \mapsto f(x))) = \alpha(y \mapsto \phi(f)) && (125) \\ &= \phi(f) \cdot \alpha(y \mapsto 1) = \phi(f) \cdot \alpha(\mathbf{1}) = \phi(f) && (126) \end{aligned}$$

A particular state α is given by $\alpha(g) = \int_Y g\, dy$ (for all $g \in L^1(Y)$). For this α the α-extension ϕ_α of ϕ is

$$\bar{\phi}(f) = \int_Y \Big(y \mapsto \phi\big(x \mapsto f(x,y)\big)\Big)\, dy \tag{127}$$

Take the σ-subalgebra $\mathcal{A} = \big\{A \times Y :\ A \in \mathcal{S}\big\}$ of $\mathcal{S} \times \mathcal{B}$ (which is isomorphic to $\mathcal{S}$). Then the $\mathcal{A}$-conditional expectation is

$$\mathbb{E}(f \mid \mathcal{A})(x,y) = \int_Y f(x,y)\, dy \tag{128}$$

We claim that for any $\alpha \in L^1(Y)^\sharp$ the state $\bar{\phi}$ is Bayes accessible from ϕ_α using the $\mathcal{A}$-conditional expectation as upgrading device; i.e. that we have

$$\bar{\phi}(f) = \phi_\alpha(\mathbb{E}(f \mid \mathcal{A})) \tag{129}$$

To show this, note first that, since the dual space $L^1(X, \mathcal{S}, p)^*$ is $L^\infty(X, \mathcal{S}, p)$ ([24][Theorem 1.7.8]), there is a function $g \in L^\infty(X)$ such that

$$\phi(f) = \int_X f(x) g(x)\, dx \quad \text{for all } f \in L^1(X) \tag{130}$$

Then for all $f \in L^1(X \times Y)$ we have

$$\begin{aligned} \bar{\phi}(f) &= \int_Y \Big(y \mapsto \phi\big(x \mapsto f(x,y)\big)\Big)\, dy && (131) \\ &= \int_Y \int_X f(x,y) g(x)\, dx dy && (132) \end{aligned}$$

Using the formula (128) giving the conditional expectation $\mathbb{E}(\cdot \mid \mathcal{A})$ and changing the order of integrals below (allowed by Fubini's theorem) we can calculate then

$$\begin{aligned} \phi\big(\mathbb{E}(f \mid \mathcal{A})\big) &= \phi\big(x \mapsto \int_Y f(x,y)\, dy\big) = \int_X \int_Y f(x,y)\, dy\; g(x)\, dx && (133) \\ &= \int_X \int_Y f(x,y) g(x)\, dy\, dx = \int_Y \int_X f(x,y) g(x)\, dx\, dy && (134) \\ &= \int_Y \Big(y \mapsto \big(\int_X f(x,y) g(x)\, dx\big)\Big)\, dy && (135) \\ &= \int_Y \Big(y \mapsto \big(x \mapsto \phi(x \mapsto f(x,y)\big)\Big)\, dy && (136) \\ &= \bar{\phi}\big(f(x,y)\big) && (137) \end{aligned}$$

Using (133) and (137) the claim (i.e. equation (129)) follows easily:

$$\begin{aligned} \phi_\alpha\big(\mathbb{E}(f \mid \mathcal{A})\big) &= \alpha\big(y \mapsto \phi(\mathbb{E}(f \mid \mathcal{A}))\big) && (138) \\ &= \alpha\big(\bar{\phi}(f)\big) && (139) \\ &= \bar{\phi}(f)\alpha(x \mapsto 1) && (140) \\ &= \bar{\phi}(f)\alpha(\mathbf{1}) = \bar{\phi}(f) && (141) \end{aligned}$$

To complete the proof one has to show that there exists an α in $L^1(Y)^\sharp$ such that $\phi_\alpha \neq \bar{\phi}$. But this is clear: take any continuous, non-constant function $t : Y \to \mathbb{R}$ for which $\int_Y t(y)\, dy = 1$ and put $\alpha(g) = \int g(y) t(y)\, dy$.

Thus we proved that ϕ has different extensions $\bar{\phi}$ and ϕ_α such that

$$\bar{\phi}(f) = \phi_\alpha\big(\mathbb{E}(f \mid \mathcal{A})\big) \tag{142}$$

∎

Proposition 8.3 shows that a Bayesian agent can learn in principle everything that can be formulated in terms of a probability measure on a fixed σ-algebra – provided the Agent has access to a potentially unlimited supply of evidence. In this sense a Bayesian Agent has unlimited learning capacity.

Note that it is *not* part of our claim that the additional evidence the Agent needs to have in order to learn a Bayes inaccessible state must be formulated in terms of the product extension of the original probability space the proof of Proposition 8.3 uses. Other extensions might very well transform a Bayes inaccessible state into a Bayes learnable one. It is even to be expected that Bayes learnability of a state via extending might depend sensitively on how the agent extends the original probability space to accommodate new knowledge.

One may wonder whether state spaces are Bayes connectable in a stronger sense than specified by Definition 8.2; i.e. whether it holds that given any pair of states ϕ and ψ in $L^1(X, \mathcal{S}, p)^\sharp$ such that ϕ is not Bayes accessible from ψ there exists an extension in which ϕ is Bayes accessible from ψ. To give the precise definition of strong Bayes connectability of state spaces, we define first the concept of in principle Bayes accessibility:

Definition 8.4. Given a state space $L^1(X, \mathcal{S}, p)^\sharp$, a state ϕ in it is called *in principle* Bayes accessible from another state $\psi \neq \phi$ if there exists a state space extension $L^1(X', \mathcal{S}', p')^\sharp$ of $L^1(X, \mathcal{S}, p)^\sharp$ such that the extension of ϕ from $L^1(X, \mathcal{S}, p)$ to $L^1(X', \mathcal{S}', p')$ is Bayes accessible from the extension of ψ from $L^1(X, \mathcal{S}, p)$ to $L^1(X', \mathcal{S}', p')$.

Definition 8.5. The state space $L^1(X, \mathcal{S}, p)^\sharp$ is called *strongly* Bayes connectable if any state ϕ is in principle Bayes accessible from any other state ψ.

Problem 8.6. Are state spaces strongly Bayes connectable?

We do not know the answer to the above question.

9 Summary and closing comments

Bayesian learning is a particular way of inferring unknown probabilities from known ones. The specificity of this kind of learning is that the inference is conditionalizing: the inferred probability measure is obtained by conditionalizing the known probability measure. We argued in this paper that conditionalizing should be carried out in terms of conditional expectations. We have seen that conditionalizing using this technique, which is standard in mathematics, includes both the elementary Bayes rule and Jeffrey conditionalization as special cases. We have shown that adopting this viewpoint leads naturally to regarding conditionalization as a two-place relation $\overset{\mathbb{E}}{\rightsquigarrow}$ in the state space determined by the reference probability measure representing the background subjective degrees of belief of a Bayesian Agent. The interpretation of $\psi \overset{\mathbb{E}}{\rightsquigarrow} \phi$ is that the Agent can learn the probabilities given by ϕ from the evidence represented by probabilities given by ψ; where "learning ϕ from ψ" means "conditionalizing ψ one obtains ϕ". Finding out the properties of the relation $\overset{\mathbb{E}}{\rightsquigarrow}$ amounts to characterizing Bayesian learning in its abstract, general form.

We have proved that the Bayes accessibility relation $\overset{\mathbb{E}}{\rightsquigarrow}$ is reflexive, antisymmetric and non-transitive. The proof of non-transitivity of Bayesian learning revealed that this feature is intimately related to the non-commutativity of upgrading probability measures via conditional expectations determined by different conditioning σ-algebras. This non-commutativity has been noticed and analyzed extensively in the literature in the special case of Jeffrey conditionalization, and it is typically found to be a conceptually very problematic feature of conditionalization via Jeffrey rule. The alleged difficulty is that non-commutativity of upgrading violates what we called in this paper the "Norm of Epistemic Commutativity": this norm requires that when an Agent draws the consequences of a whole set of information, it should not matter in what order the elements of the information set are presented to the Agent. We argued that interpreting non-commutativity of conditionalization via conditional expectation as violation of the Norm of Epistemic Commutativity is not justified. A technically explicit specification of the informal Norm of Epistemic Commutativity in terms of conditional expectations shows that the Norm of Epistemic Commutativity is in fact satisfied by upgrading via conditional expectations.

We also have investigated the connectivity properties of state spaces with respect to the Bayes accessibility relation $\overset{\mathbb{E}}{\rightsquigarrow}$. We have shown that state spaces are typically not weakly Bayes connected. That is to say, we proved that, given a measure representing the background degrees of belief of a Bayesian Agent, there exist states (probability measures) that cannot be learned by the Agent from any evidence the Agent is capable of formulating within the confines of a given probability measure space. Unlike failure of transitivity of the Bayes accessibility relation, failure of weak Bayes connectedness seems to pose a very serious challenge for Bayesian learning: The existence of Bayes inaccessible states (we have proved that there exist an uncountably infinite number of such states in the typical cases) means that an Agent's background measure might prohibit the Agent from learning the "true" probability measure. If the true probability measure happens to be one of the Bayes inaccessible ones, the Agent cannot learn it by conditionalizing. Thus the Agent's background knowledge proves to be crucial from the perspective of what the Agent can in principle learn from possible evidence. In particular, the state spaces of standard probability measure spaces are not weakly Bayes connected. This is a very large class that includes most applications. In these probability theories the true probability measure to be learned might remain inaccessible for the Bayesian Agent.

A Bayesian Agent might try to overcome the epistemological difficulty posed by Bayes inaccessible states by widening the probabilistic framework in which Bayes inaccessible states are present. This strategy involves enlarging the σ-algebra of propositions stating features of the world, extending the background probability to the enlarged set, and looking for evidence about (probabilities of) some subset of the enlarged Boolean algebra – all this in the hope of becoming able to Bayes-learn those probabilities in the broader framework that are inaccessible in a narrower probability theory. We showed that such a strategy is in principle viable: a Bayes inaccessible state becomes Bayes learnable from some state after a suitable embedding of the original probability space into a larger one. Thus, a Bayesian Agent has unlimited learning capacity if he is allowed to expand the propositional base of possible evidence. It is even possible to enlarge the probability space into one in which Bayes inaccessible states do not exist and thus every probability is Bayes learnable from some evidence: We showed that state spaces of large enough probability spaces are weakly Bayes connected. A Bayesian Agent can only do such an extension however if he is capable of comprehending a very large amount of propositions: The σ-algebra of the probability space which could be shown having a Bayes connected state space had cardinality larger than the cardinality of the set of Lebesgue measurable subsets of real numbers. Since the cardinality of the set of Lebesgue measurable sets itself is already beyond the continuum, one needs an extremely strong concept of Bayesian Agent to allow for this option. A Bayesian Agent with a more modest mental capacity has to be aware however that he is on an unended quest: for him in every probability space he is able to comprehend there exist probability statements that might be true but he only can learn them from evidence that can be gathered only by going beyond the framework in which the true probability is formulated. Whether the concept of a powerful Bayesian Agent is reasonable, and whether the notion of a modest Bayesian Agent is attractive, we do not wish to try to decide here.

Acknowledgement

Research supported in part by the Hungarian Scientific Research Found (OTKA, contract numbers: K100715 and K115593). M. Rédei thanks the Munich Center for Mathematical Philosophy in the Ludwig-Maximilians University (Munich, Germany), with which he was affiliated as Visiting Fellow, and the Institute of Philosophy of the Hungarian Academy of Sciences, where he was an Honorary Research Fellow, while this paper was written.

References

[1] P. Billingsley. *Probability and Measure*. John Wiley & Sons, New York, Chichester, Brisbane, Toronto, Singapore, Third edition, 1995.

[2] V.I. Bogachev. *Measure Theory*, volume II. Springer, Berlin, Heidelberg, New York, 2007.

[3] L. Bovens and S. Hartmann. *Bayesian Epistemology*. Oxford University Press, 2004.

[4] R. Bradley. Radical probabilism and Bayesian conditioning. *Philosophy of Science*, 72:342–364, 2005.

[5] I. Csiszar. I-divergence geometry of probability distributions of minimization problems. *Annals of Probability*, 3:146–158, 1975.

[6] P. Diaconis and S.L. Zabell. Updating subjective probability. *Journal of the American Statistical Association*, 77:822–830, 1982.

[7] Z. Domotor. Probability kinematics and representation of belief change. *Philosophy of Science*, 47:384–403, 1980.

[8] J. Doob. The development of rigor in mathematical probability theory (1900-1950). *American Mathematical Monthly*, pages 586–595, 1996.

[9] J.L. Doob. *Stochastic processes*. John Wiley & Sons, 1953.

[10] F. Döring. Why Bayesian psychology is incomplete. *Philosophy of Science*, 66:S379–S389, 1999.

[11] R.G. Douglas. Contractive projections on an L_1-space. *Pacific Journal of Mathematics*, pages 443–462, 1965.

[12] W. Feller. *An Introduction to Probability Theory and its Applications*, volume 2. Wiley, New York, 2nd edition, 1971. First edition: 1966.

[13] H. Field. A note on Jeffrey conditionalization. *Philosophy of Science*, 45:361–367, 1978.

[14] D.H. Fremlin. *Measure Theory*, volume 2. Torres Fremlin, 2001.

[15] J.D. Gallow. How to learn from theory-dependent evidence; or commutativity and holism: A solution for conditionalizers. *The British Journal for the Philosophy of Science*, 65:493–519, 2014.

[16] D. Garber. Field and Jeffrey conditionalization. *Philosophy of Science*, 47:142–145, 1980.

[17] B. Gyenis. Bayes rules all. Submitted, 2015.

[18] P. Halmos and S. Givant. *Logic as Algebra.* Number 21 in Dolciani Mathematical Expositions. The Mathematical Association of America, 1998.

[19] C. Howson. Bayesian rules of updating. *Erkenntnis*, 45:195–208, 1996.

[20] C. Howson. Finite additivity, another lottery paradox, and conditionalization. *Synthese*, 191:989–1012, 2014.

[21] C. Howson and A. Franklin. Bayesian conditionalization and probability kinematics. *The British Journal for the Philosophy of Science*, 45:451–466, 1994.

[22] C. Howson and P. Urbach. *Scientific Reasoning: The Bayesian Approach.* Open Court, Illinois, 1989. Second edition: 1993.

[23] R.C. Jeffrey. *The Logic of Decision.* The University of Chicago Press, Chicago, first edition, 1965.

[24] R.V. Kadison and J.R. Ringrose. *Fundamentals of the Theory of Operator Algebras*, volume I. and II. Academic Press, Orlando, 1986.

[25] G. Kalmbach. *Orthomodular Lattices.* Academic Press, London, 1983.

[26] A.A. Kirillov and A.D. Gvishiani. *Theorems and Problems in Functional Analysis.* Problem Books in Mathematics. Springer-Verlag, New York, 1982.

[27] A.N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Springer, Berlin, 1933. English translation: Foundations of the Theory of Probability, (Chelsea, New York, 1956).

[28] M. Lange. Is Jeffrey conditionalization defective in virtue of being noncommutative? Remarks on the sameness of sensory experience. *Synthese*, 123:393–403, 2000.

[29] M. Loéve. *Probability Theory.* D. Van Nostrand, Princeton, Toronto, London, Melbourne, 3rd edition, 1963.

[30] K. Petersen. *Ergodic Theory.* Cambridge University Press, Cambridge, 1989.

[31] J. Pfanzagl. Characterizations of conditional expectations. *The Annals of Mathematical Statistics*, 38:415–421, 1967.

[32] M.M. Rao. *Conditional Measures and Applications.* Chapman & Hall/CRC, Boca Raton, London, New York, Singapore, 2nd, revised and expanded edition, 2005.

[33] M. Rédei. *Quantum Logic in Algebraic Approach*, volume 91 of *Fundamental Theories of Physics.* Kluwer Academic Publisher, 1998.

[34] J.S. Rosenthal. *A First Look at Rigorous Probability Theory.* World Scientific, Singapore, 2006.

[35] W. Rudin. *Real and Complex Analysis.* McGraw-Hill, Singapore, 3rd edition, 1987.

[36] B. Skyrms. *Choice and Chance.* Wadsworth Publishing Co., Belmont, 3rd edition, 1986.

[37] B.C. van Fraassen. *Laws and Symmetry.* Claredon Press, Oxford, 1989.

[38] A. Villani. Another note on the inclusion $L^p(\mu) \subset L^q(\mu)$. *The American Mathematical Monthly*, 92:485–487, 1985.

[39] C. Wagner. Probability kinematics and commutativity. *Philosophy of Science*, 69:266–278, 2002.

[40] C. Wagner. Is conditioning really incompatible with holism? *Journal of Philosophical Logic*, 42:409–414, 2013.

[41] J. Weisberg. Commutativity or holism? A dilemma for conditionalizers. *The British Journal for the Philosophy of Science*, 60:793–812, 2009.

[42] J. Weisberg. Updating, undermining, and independence. *The British Journal for the Philosophy of Science*, 66:121–159, 2015.

[43] J. Weisberg. You've come a long way, Bayesians. *Journal of Philosophical Logic*, 2015. 40th Anniversary Issue, forthcoming.

[44] J. Williamson. *In Defence of Objective Bayesianism*. Oxford University Press, Oxford, 2010.