The Bayes Blind Spot of a finite Bayesian Agent is a large set*

Zalán Gyenis† Miklós Rédei‡

August 1, 2016

Abstract

The Bayes Blind Spot of a Bayesian Agent is the set of probability measures on a Boolean algebra that are absolutely continuous with respect to the background probability measure (prior) of a Bayesian Agent on the algebra and which the Bayesian Agent cannot learn by conditionalizing no matter what (possibly uncertain) evidence he has about the elements in the Boolean algebra. It is shown that if the Boolean algebra is finite, then the Bayes Blind Spot is a very large set: it has the same cardinality as the set of all probability measures (continuum); it has the same measure as the measure of the set of all probability measures (in the natural measure on the set of measures); and is a "fat" (second Baire category) set in topological sense in the set of all probability measures taken with its natural topology.

1 Learning by conditionalizing

A Bayesian Agent is an abstract, ideal person having degrees of belief $p(C)$ about (the truths of) propositions C in a set $\mathcal{S}$ forming a Boolean algebra. The degrees of belief $p(C)$ behave like probabilities: p is an additive map on $\mathcal{S}$ formed by (some) subsets of the set X of elementary propositions. The triplet $(X, \mathcal{S}, p)$ is a probability measure space [1], [11]. Throughout this paper it is assumed that the Boolean algebra $\mathcal{S}$ has a finite number of elements. (In section 4 we will comment on the situation when $\mathcal{S}$ is infinite.)

A Bayesian Agent is able to learn: Suppose the Agent is told that proposition $A \in \mathcal{S}$ is true (but nothing else about other propositions in $\mathcal{S}$). Using his background probability p, if $p(A) \neq 0$, the Agent can infer from this information probabilities $q(B)$ of events B other than A by conditionalizing p via A using Bayes' rule:

$$q(B) = \frac{p(B \cap A)}{p(A)} \qquad \text{for all } B \in \mathcal{S} \tag{1}$$

q is a new probability measure on $\mathcal{S}$; it can be viewed as the probability measure that the Agent has inferred, on the basis of his prior p, from the probability measure $q_{\mathcal{A}}$ that is defined on the

*Research supported in part by the Hungarian Scientific Research Found (OTKA). Contract numbers: K-115593 and K100715.

†BUTE Department of Algebra, Budapest, Hungary, gyz@renyi.hu. Work done while staying in the Centre for the Philosophy of Natural and Social Sciences of London School of Economics and Political Science on a grant awarded by the European Philosophy of Science Association.

‡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

four element Boolean subalgebra $\{\emptyset, A, A^\perp, X\}$ of $\mathcal{S}$ that is generated by A and $A^\perp$ and which has the feature that it takes values $q_{\mathcal{A}}(A) = 1$ and $q_{\mathcal{A}}(A^\perp) = 0$ on the non-trivial elements of $\mathcal{A}$. The probability measure $q_{\mathcal{A}}$ represents *certain evidence* [7][p. 452]. Note that q has value 0 on every element B which has p-probability zero. The technical expression of this feature of q is that q is absolutely continuous with respect to p [1][p. 422].

Suppose that the Agent receives information about A and $A^\perp$ that is given by a probability measure $q_{\mathcal{A}}$ which does not have the extreme values 1 and 0 but the values $q_{\mathcal{A}}(A) = r \neq 1$ and $q_{\mathcal{A}}(A^\perp) = 1 - r \neq 0$. What probability measure can the Agent infer from this evidence on the basis of the background measure p? The standard answer to this question is: If neither $p(A)$ nor $p(A^\perp)$ is equal to zero, then the Agent can use the Jeffrey conditionalization rule [8] to obtain the measure q:

$$q(B) \doteq \frac{p(B \cap A)}{p(A)} q_{\mathcal{A}}(A) + \frac{p(B \cap A^\perp)}{p(A^\perp)} q_{\mathcal{A}}(A^\perp) \qquad \text{for all } B \in \mathcal{S} \tag{2}$$

More generally, if the evidence the Agent has are the probabilities $q_{\mathcal{A}}(A_i)$ of mutually disjoint events A_i $(i = 1, 2 \ldots N)$ forming a *non-trivial* partition in $\mathcal{S}$, which generates the *proper* Boolean subalgebra $\mathcal{A}$ of $\mathcal{S}$, and if these events have non-zero prior probability $p(A_i) \neq 0$ (for all i), then the Agent can infer from this so-called *uncertain evidence* [2], [15] a probability measure q using the general Jeffrey conditionalizing rule:

$$q(B) \doteq \sum_i \frac{p(B \cap A_i)}{p(A_i)} q_{\mathcal{A}}(A_i) \qquad \text{for all } B \in \mathcal{S} \tag{3}$$

Just like in the case of conditionalization via Bayes' rule, q obtained this way is absolutely continuous with respect to the prior probability p. To simplify matters, from now on we assume that the prior probability of the Agent is non-zero on every element $\{x\}$ for $x \in X$. In this case, obviously, every probability measure on $\mathcal{S}$ is absolutely continuous with respect to p (see Remark 3.5 for general prior probability). Note that the requirement that the uncertain evidence is given by a probability measure on a *non-trivial* partition (equivalently: on a *proper* Boolean subalgebra $\mathcal{S}$) is important: if the evidence were taken to be a probability measure q' on the whole $\mathcal{S}$, then for every element x in X the Jeffrey rule (3) would entail

$$q(\{x\}) = \sum_i \frac{p(\{x\} \cap A_i)}{p(A_i)} q'(A_i) = \frac{p(\{x\})}{p(\{x\})} q'(\{x\}) = q'(\{x\}) \tag{4}$$

This equation says that every probability measure can be obtained from itself as evidence via the Jeffrey rule – a triviality.

As an elementary example for the Jeffrey conditionalization consider die throwing: Let $X_6 = \{x_1, x_2, \ldots, x_6\}$ represent the possible outcomes of throwing a die, and let $\mathcal{S}_6$ be the Boolean algebra of subsets of X_6. Assume that the Agent's background probability p is given on elements $x \in X_6$ according to Figure 1 below.

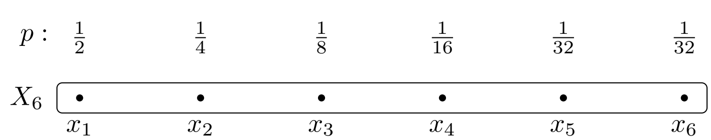


Figure 1: Example of probabilities in die throwing

Consider the partition

$$A_1 = \{x_1, x_2\} \qquad A_2 = \{x_3, x_4, x_5\} \qquad A_3 = \{x_6\} \tag{5}$$

indicated in Figure 2. Suppose the Agent receives the information $q_{\mathcal{A}}$, where the probability measure $q_{\mathcal{A}}$ is given on the elements of the partition A_1, A_2, A_3 by

$$q_{\mathcal{A}}(A_1) = \frac{2}{6} \qquad q_{\mathcal{A}}(A_2) = \frac{3}{6} \qquad q_{\mathcal{A}}(A_3) = \frac{1}{6} \tag{6}$$

Using the Jeffrey conditionalization rule (3), the Agent can infer from evidence $q_{\mathcal{A}}$ the probability measure q indicated in Figure 2:

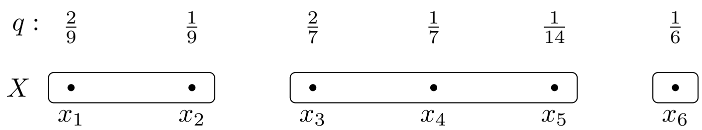


Figure 2: Example of inferring probabilities using Jeffrey conditionalization

2 The Bayes Blind Spot

Consider now the question: Suppose the true probability distribution describing the results of throws with a die is q given by the Figure 3 below. Can the Bayesian Agent (having p as his background measure) infer this probability q from *some* probability measure as evidence by conditionalizing using the Jeffrey rule (3)? If so, we call q *Bayes accessible* or *Bayes learnable*.

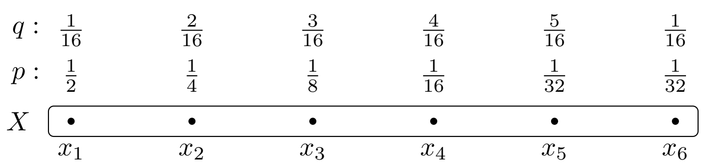


Figure 3: Is q Bayes accessible?

The question whether q is Bayes accessible is asking whether there exists a *non-trivial* partition of the 6 element set X_6 and a probability measure $q_{\mathcal{A}}$ defined on elements of this partition such that q can be obtained from $q_{\mathcal{A}}$ in the manner (3). The question is not trivial: there exist 203 different partitions in $\mathcal{S}$ (203 is the 6$^{\text{th}}$ Bell number [4][p. 91-93). Thus, if one would try to answer the question by "brute force", one would have to consider all the 203 partitions and, for each partition, write out eq. (3) for every B to obtain a large number of equations to solve with $q_{\mathcal{A}}(A_i)$ as unknowns to see if the system of equations admit a solution. While doable, this procedure becomes intractable in the general situation when the number of the elements in the Boolean algebra is very large. One can however find a simple, compact condition that can be used to decide whether a probability measure can be obtained as a conditional probability via the Jeffrey conditionalization:

Suppose we have found a partition $\{A_i\}$ and a $q_{\mathcal{A}}$ for which q can be written in the form (3). If the partition $\{A_i\}$ is non-trivial, then at least one of A_i has more than one elements from X_6. Suppose A_i has two elements x_1 and x_2. Then (3) entails

$$q(\{x_1\}) = \sum_i \frac{p(\{x_1\} \cap A_i)}{p(A_i)} q_{\mathcal{A}}(A_i) = \frac{p(\{x_1\})}{p(A_i)} q_{\mathcal{A}}(A_i) \tag{7}$$

$$q(\{x_2\}) = \sum_i \frac{p(\{x_2\} \cap A_i)}{p(A_i)} q_{\mathcal{A}}(A_i) = \frac{p(\{x_2\})}{p(A_i)} q_{\mathcal{A}}(A_i) \tag{8}$$

Equations (7)-(8) entail that a necessary condition for q to be Bayes accessible is that the following condition holds:

$$\frac{q(\{x_1\})}{p(\{x_1\})} = \frac{q(\{x_2\})}{p(\{x_2\})} \tag{9}$$

One can verify easily that the probability measure q describing the distribution of throws with a die with values indicated in Figure 3 *violates* condition (9). Consequently, this probability measure is *not* Bayes accessible: A Bayesian Agent having his background knowledge represented by the probability measure p given in Figure 1 is not able to learn this q distribution via conditionalizing no matter what (possibly uncertain) evidence he is presented with.

The reasoning leading to the necessary condition (9) for Bayes accessibility generalizes easily from $\mathcal{S}_6$ to an arbitrary finite Boolean algebra. This, in turn leads to a sufficient condition entailing that a probability measure is *not* Bayes accessible: If for a probability measure q on $\mathcal{S}$ we have

$$\frac{q(\{x_i\})}{p(\{x_i\})} \neq \frac{q(\{x_j\})}{p(\{x_j\})} \qquad i \neq j;\ 1 \leq i, j \leq n \tag{10}$$

then q is *not* Bayes accessible for the Bayesian Agent having p as his background degree of belief.

The function $\frac{dq}{dp}$ defined by

$$X \ni x_i \mapsto \frac{dq}{dp}(x_i) \doteq \frac{q(\{x_i\})}{p(\{x_i\})} \tag{11}$$

is known as the Radon-Nikodym derivative (also called the density) of q with respect to p [1][p. 423]. Thus, the content of the necessary condition (10) can be expressed compactly by saying that q is not Bayes accessible for the Bayesian Agent having background probability p if the Radon-Nikodym derivative $\frac{dq}{dp}$ of q with respect to p is an injective function. We show now that this condition also is necessary, i.e. we will prove

Proposition 2.1 (cf. [6])**.** Let $(X, \mathcal{S}, p)$ be a probability space with a finite set X having n elements and $\mathcal{S}$ the Boolean algebra of subsets of X. A probability measure q on $\mathcal{S}$ is *not* Bayes accessible if and only if its Radon-Nikodym derivative $\frac{dq}{dp}$ is an injective function.

Proof. Since we have seen that injectivity of the Radon-Nikodym derivative is sufficient for Bayes inaccessibility, we only have to show that injectivity is also necessary, i.e. that non-injectivity entails Bayes accessibility. Let the range of $\frac{dq}{dp}$ be $\{y_1, \ldots, y_k\}$. If $\frac{dq}{dp}$ is not injective, then the partition

$$A_i = \left\{x \in X : \ \frac{dq}{dp}(x) = y_i\right\} \quad \text{for } i = 1 \ldots k$$

is a non-trivial partition of X i.e. there is at least one A_i containing at least two elements. Note that $\frac{dq}{dp}$ is constant on every A_i. We define the probability measure r on the Boolean subalgebra generated by the partition A_i by defining the values of r on the blocks of the partition and requiring r to be additive:

$$r(A_i) = \frac{q(\{x\})}{p(\{x\})} p(A_i) \qquad \text{for any } x \in A_i$$

Then, for all $x \in X$ there is a unique j such that $x \in A_j$ and thus we have

$$\sum_i \frac{p\big(\{x\} \cap A_i\big)}{p(A_i)} r(A_i) = \frac{p(\{x\})}{p(A_j)} r(A_j) = q(\{x\})$$

∎

As the example of die throwing shows, Bayes inaccessible states can exist. More generally, one can show that given any background probability p on a finite Boolean algebra, there exists a q on that Boolean algebra that is Bayes inaccessible [6]. Following the terminology introduced in [6] we will call the set of probability measures on $\mathcal{S}$ that are not Bayes accessible for the Bayesian Agent (with respect to the fixed background probability p) the "Bayes Blind Spot" of the Agent. If the p-dependence of the Bayes Blind Spot needs to be made explicit, we say "Bayes p-Blind Spot".

3 Size of the Bayes Blind Spot

How large is the Bayes Blind Spot? There is no unique answer to this question: The size of a set can be gauged using conceptually different "yardsticks". Given a yardstick, one can compare the size of a set to the sizes of other sets, measured by the same yardstick. There are three standard ways to measure the size of a set [12][p. 170] and thus also the size of the Bayes Blind Spot:

Cardinality One can ask what the cardinality of the Bayes Blind Spot is and how its cardinality is related to the cardinality of the set of all probability measures.

Topological size One can ask whether the Bayes Blind Spot is a meager (Baire first category) or nonmeager (Baire second category) set in the set of all probability measures with respect to a natural topology.

Measure theoretical size One can ask what the size of the Bayes Blind Spot is with respect to a measure on the set of all probability measures.

We show now that the Bayes Blind Spot is a *very large* set in the sense of all the three measures – cardinality, topological and measure theoretical size.

3.1 Cardinality

The sufficient condition (10) for Bayes inaccessibility makes it clear that if q' is Bayes inaccessible, then for all small enough positive real numbers ε the probability measures q_ε such that

$$|q_\varepsilon(\{x\}) - q'(\{x\})| \leq \varepsilon \qquad \text{for all } x \in X \tag{12}$$

also satisfy (10) and thus are not Bayes accessible. It follows from this that the Bayes Blind Spot has at least continuum cardinality [6]. On the other hand, the cardinality of the set of *all* probability measures on a finite Boolean algebra is at most the continuum: a probability measure is a function from the finite set X having n elements into the unit interval $[0, 1]$; so the set of all probability measures on X is a subset of the cartesian product $\times_1^n[0, 1]$, cardinality of which is the same as the cardinality of $[0, 1]$. It follows that we have the following

Proposition 3.1. The Bayes Blind Spot of a Bayesian Agent has the cardinality of the continuum, and, consequently, for a Bayesian Agent there exist exactly as many Bayes inaccessible probability measures as the number of all probability measures (in the sense of cardinality), namely a continuum number.

3.2 Topological size – Baire category

Recall that, given a subset E of a topological space T, point x in T is an interior point of E if there is an open set O such that x is belongs to O and O is contained in E. The set of all interior points of E is called the interior of E. A subset E of T is said to be nowhere dense if its closure has empty interior. The sets of the first Baire category in T are those that are countable unions of nowhere dense sets [13][p. 42]. Any subset of T that is not of the Baire first category is said to be of the second Baire category. A set E is nowhere dense if and only if its complement $T \setminus E$ contains an open set that is dense in T. Thus a subset of T which is open and dense is of the second Baire category.

Sets of first category are "meager", whereas sets of second category are regarded as nonmeager ("fat") in a topological sense. To see why, it is useful to have examples.

Consider the real line $\mathbb{R}$ with its usual topology. Any finite set of points on the line is a nowhere dense set. The set $\mathbb{Q}$ of rational numbers is a meager set because $\mathbb{Q}$ is a countable union of single rational numbers.

Non-countable meager sets also exists: the Cantor set is uncountable, closed, compact and nowhere dense in $\mathbb{R}$ (see [14]). The Cantor set is large in cardinality (within the set of real numbers), small in the sense of topology and also small measure theoretically: it is a null-set with respect to the Lebesque measure. But a meager set can have large measure: the real line can be decomposed into two disjoint sets, one being of first Baire category, the other having measure zero with respect to the Lebesgue measure (Theorem 1.6 in [9]). Such a set is the fat Cantor set, [14], which is meager but can have arbitrary large measure.

Open dense sets are easy to come up with: obviously $\mathbb{R}$ is open and dense in itself. Removing a finite number of points from $\mathbb{R}$ one obtains an open dense set. Less obvious example is the complement of the Cantor set: since the Cantor set is closed and nowhere dense, its complement is open and dense.

To assess the topological size of the Bayes Blind Spot in the set $M(\mathcal{S})$ of all probability measures on $\mathcal{S}$, we need to specify a topology on $M(\mathcal{S})$. Topologies can be defined by metrics (distance functions), and this is how one can specify a topology in the set of probability measures. There exist several types of metrics among probability measures that one can consider. The Appendix lists five typical ones that occur in different contexts. It turns out (and this is proved in the Appendix) that they all are equivalent in the sense that they determine the same topology, which we will call *the standard uniform topology.* The content of this topology can be expressed in different ways, one of which is the formulation in terms of the distance d_3 of the Appendix: if the probability measure q is d_3-close to the probability measure q' then the supremum of the difference of the expectation values of random variables with respect to q and q' is small among *all* the random variables whose expectation values with respect to the background probability p are close.

Given the standard uniform topology, the topological size of the Bayes Blind Spot is the characterized by the following proposition (proof of which we give in the Appendix):

Proposition 3.2. The Bayes Blind Spot is an open and dense set in the set $M(\mathcal{S})$ of all probability measures equipped with the standard uniform topology on the probability measures.

Corollary 3.3. The complement of the Bayes Blind Spot, the set of Bayes accessible probability measures is a closed, nowhere dense set in the standard uniform topology on the probability measures.

Proposition 3.2 says that the Bayes Blind Spot is a very large, a "fat" set in topological sense, much larger than the set of Bayes accessible states. Viewed from the perspective of topology, there exist much more Bayes inaccessible states than Bayes accessible ones.

Corollary 3.3 entails that the limit of Bayes accessible probability measures is again Bayes accessible. Consequently, a Bayes inaccessible probability measure cannot be approximated with arbitrary precision by Bayes accessible probability measures. Thus one cannot "neutralize" the presence of Bayes inaccessible states by taking the position that the Bayesian Agent can in principle be presented with a series of evidences that can get him arbitrarily close to a Bayes inaccessible probability measure.

Furthermore, the set of Bayes accessible probability measures, being the complement of a dense open set, is not only a closed set but a *meager* set: a closed set with empty interior. Thus, while there exist an uncountable infinite number of Bayes inaccessible probability measures arbitrary close to every Bayes accessible one, every Bayes inaccessible probability measure has a neighborhood in which there are *only* Bayes inaccessible probability measures.

The Bayes inaccessible probability measures "dominate" the set of all probability measures completely in a topological sense.

3.3 Measure theoretical size

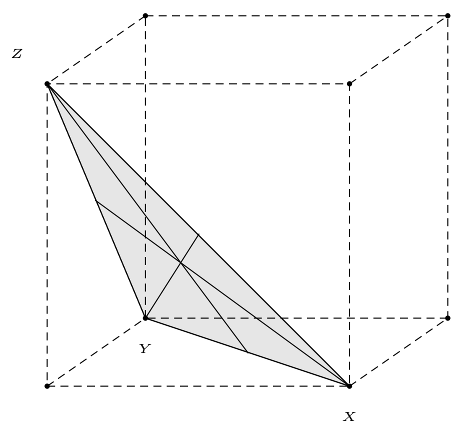
To assess the measure theoretical size of the Bayes Blind Spot in the set $M(\mathcal{S})$, one has to specify a σ-algebra in $M(\mathcal{S})$ and a measure over this algebra. The natural algebra and measure is the one arising from the Lebesgue measure in the following way:

We can identify measures in $M(\mathcal{S})$ with functions $f : X \to [0,1]$ such that $\sum_{x\in X} f(x) = 1$. Under this identification each probability measure is identified with a point in $[0,1]^n$ (recall: n is the number of elements in X). Thus $M(\mathcal{S}) \subseteq [0,1]^n$.

The equation

$$X_1 + X_2 + \ldots + X_n = 1 \qquad (X_i \in \mathbb{R} \text{ a variable}) \tag{13}$$

defines an $n-1$-dimensional hyperplane H in $\mathbb{R}^n$; thus $M(\mathcal{S})$ is the simplex which is the intersection of this hyperplane with the unit cube $[0,1]^n$ (see the picture below).



For any finite dimension d the d-dimensional Lebesgue measure λ^d is defined on the Borel sets of the d-cube $[0,1]^d$. Since $M(\mathcal{S}) \subseteq H$ is a subset of an $n-1$ dimensional hyperplane in $\mathbb{R}^n$, we have $\lambda^n(M(\mathcal{S})) = 0$. On the other hand with λ^{n-1} being the Lebesgue measure on the Borel sets of $H \cap [0,1]^n$ we have

$$\lambda^{n-1}\big(M(\mathcal{S})\big) = \lambda^{n-1}\big(H \cap [0,1]^n\big) > 0 \tag{14}$$

The measure

$$\mu \equiv \frac{\lambda^{n-1}}{\lambda^{n-1}(M(\mathcal{S}))} \tag{15}$$

is the normalized area (Lebesgue) measure on $M(\mathcal{S})$; in this measure the whole set $M(\mathcal{S})$ of probability measures has measure equal to 1. The next proposition (proved in the Appendix) states the size of the Bayes Blind Spot in this measure.

Proposition 3.4. The Bayes Blind Spot has μ measure equal to 1. The set of Bayes accessible states is a μ measure zero set.

Proposition 3.4 says that the Bayes Blind Spot is a *very large* set in the set of all probability measures, with respect to the natural (Lebesgue) measure in which the set of all probability measures has non-zero measure. "Very large" means here: as large as possible: having the same size as the size of the set of *all* probability measures. This entails that the Bayes accessible states form a measure zero set in this measure.

Remark 3.5. Propositions 3.2, 3.1 and 3.4 are proved under the assumption that the background probability measure p is faithful. These propositions remain true however if the faithfulness assumption is dropped: If p is not a faithful probability measure, then it has zero probability on some elements in X. In terms of the geometrical picture of Figure 3.3 this means that the point in the simplex representing p is on an "edge" E of the simplex. All the probability measures that are absolutely continuous with respect to p, hence all the potentially Bayes p-accessible probability measures, are also on E. This edge can be regarded as the set of all probability measures on the Boolean algebra that is obtained from $\mathcal{S}$ by removing from $\mathcal{S}$ the one-element sets on which p is zero, and the restriction p' of p to this Boolean algebra is faithful. Proposition 3.2 entails then, that the set of Bayes p'-accessible probability measures is a nowhere dense set in E in the relative topology on E inherited from $M(\mathcal{S})$. But then this set also is a nowhere dense set in $M(\mathcal{S})$, and its complement, the Bayes p'-Blind Spot, contains an open dense set, and is thus a set of Baire second category. It follows that the Bayes p-Blind Spot is a set of second Baire category, irrespective of wether p is faithful or not. Since an open and dense set in a complete metric space has to have uncountable cardinality, the Bayes p-Blind Spot has uncountable cardinality irrespective of wether p is faithful or not. Furthermore, since the edge E lies in a proper linear subspace of the linear space in which $M(\mathcal{S})$ has non-zero λ^{n-1} (Lebesgue) measure, the measure of the set of Bayes p'-accessible measures in E also has λ^{n-1} measure zero. It follows that the Bayes p-Blind spot has measure 1 in the measure μ in which $M(\mathcal{S})$ has measure 1 too – irrespective of whether p is faithful.

4 Concluding remarks

The notion of Bayes Blind Spot can also be defined for probability measure spaces $(X, \mathcal{S}, p)$ with an infinite Boolean algebra $\mathcal{S}$. In this more general situation the general conditioning rule yielding conditional probabilities with respect to arbitrary Boolean σ algebras of $\mathcal{S}$ is given by the concept of conditional expectation, of which the Jeffrey rule is just a particular case [6]. Determining the size of the Bayes p-Blind Spot of a general probability measure space $(X, \mathcal{S}, p)$ is a non-trivial problem, with a number of questions still open. The following partial results are known in the general case:

One can give an abstract, general characterization of probability spaces with non-empty Bayes Blind Spot [6]. On the basis of that characterization one can show the following:

- There exist probability spaces with an empty Bayes Blind Spot. The only example of such a probability space known to us is the one constructed in [6]. The set of elementary events X of this probability space is very large: its cardinality $|X|$ has to satisfy $|X| > 2^{2^{\aleph_0}}$ (with $\aleph_0$ being the countable cardinality).
- The "usual" (technically speaking: the "standard", see Definition 4.5 in [10]) infinite probability spaces that occur in applications can be shown to have a Bayes Blind Spot that has the cardinality of the continuum [6]. Such probability spaces include the probability measures on $\mathbb{R}^n$ given by a density function with respect to the Lebesgue measure in $\mathbb{R}^n$. Work is in progress to determine the topological and measure theoretical size of the Bayes Blind Spot of these standard

probability spaces [5].

The presence of large Bayes Blind Spots indicates the crucial importance of the prior probability in Bayesian learning based on a single act of conditionalization. The limits of what can be learned in a single probabilistic inference on the basis of a prior are extremely restrictive in case of a probability theory with a finite set of random events. This leads naturally to the question of whether repeated learning can make a probability measure in the Bayes Blind Spot Bayes learnable in more than one step. To investigate this question requires defining Bayesian learning dynamics precisely. It turns out that Bayes dynamics can be specified in a number of non-equivalent ways. A preliminary exploration of the behavior of Bayes dynamical systems indicate that the features of different Bayes dynamics from the perspective of their Bayes Blind Spots differ significantly. A paper is in preparation that investigates the Bayes Blind Spots of general Bayesian dynamical systems [5].

Appendix

4.1 Metrics and topology in the set of probability measures on a finite Boolean algebra

Definition 4.1. For $q, r \in M(\mathcal{S})$ we define the following metrics.

Chebysev distance:

$$d_0(q,r) = \max_{x\in X} |q(\{x\}) - r(\{x\})| \tag{16}$$

Total variation distance I:

$$d_1(q,r) = \frac{1}{2}\sum_{x\in X} |q(\{x\}) - r(\{x\})| \tag{17}$$

Total variation distance II:

$$\begin{aligned} d_2(q,r) &= \sup\left\{\left|\sum_{x\in X} f(x)q(\{x\}) - \sum_{x\in X} f(x)r(\{x\})\right| :\ f = \chi_E,\ E\in\mathcal{S}\right\} \\ &= \max_{E\in\mathcal{S}} |q(E) - r(E)| \end{aligned}$$

$\|\cdot\|_1$-distance:

$$d_3(q,r) = \sup\left\{\left|\sum_{x\in X} f(x)q(\{x\}) - \sum_{x\in X} f(x)r(\{x\})\right| :\ f\in L^1(X,\mathcal{S},p),\ \|f\|_1 \leq 1\right\} \tag{18}$$

$\|\cdot\|_\infty$-distance of density functions:

$$d_4(q,r) = \sup_{x\in X}\left|\frac{dq}{dp}(x) - \frac{dr}{dp}(x)\right| \tag{19}$$

Hellinger distance:

$$d_5(q,r) = \sum_{x\in X}\left(\left(\sqrt{\frac{dq}{dp}(x)} - \sqrt{\frac{dr}{dp}(x)}\right)^2 \cdot p(x)\right) \tag{20}$$

Euclidean distance:

$$d_6(q,r) = \left(\sum_{x\in X} (q(\{x\}) - r(\{x\}))^2\right)^{\frac{1}{2}} \tag{21}$$

Two metrics d and d' on a set M are said to be equivalent if there are constants A and B such that

$$A \cdot d(x,y) \le d'(x,y) \le B \cdot d(x,y) \qquad \text{for all } x,y \in M$$

Equivalent metrics generate the same topology on M (see [3][p. 121]).

Proposition 4.2. d_i generate the same topology on $M(\mathcal{S})$ for all $i = 0 \dots 6$.

Proof. It is straightforward to check $d_0 \le 2d_1 \le |X|d_0$. That d_0 and d_6 are equivalent follows from $d_0 \le d_6 \le 2d_1$. Next, we show $d_1 = d_2$. Let $A = \{x \in X : \ q(\{x\}) \ge r(\{x\})\}$. Then

$$\begin{aligned} d_1(q,r) &= \frac{1}{2}\sum_{x\in X} |q(\{x\}) - r(\{x\})| \\ &= \frac{1}{2}\left(\sum_{x\in A} q(\{x\}) - r(\{x\}) + \sum_{x\in X\setminus A} r(\{x\}) - q(\{x\})\right) \\ &= \frac{1}{2}\left(q(A) - r(A) + r(X\setminus A) - q(X\setminus A)\right) = q(A) - r(A) \\ &= \max_{E\subseteq X} |q(E) - r(E)| = d_2(q,r). \end{aligned}$$

That d_1 and d_5 are equivalent follows from the Cauchy–Schwartz inequality:

$$d_1(q,r) \le 2d_5(q,r) \le 2d_1(q,r)^{1/2}$$

Next, we claim $d_3 = d_4$. Any probability measure q on $\mathcal{S}$ defines a linear functional ϕ_q on $L^1(X,\mathcal{S},p)$ by assigning to any $f \in L^1(X,\mathcal{S},p)$ its expectation value with respect to q:

$$\phi_q(f) = \int_X f dq = \sum_{x\in X} f(x)q(\{x\})$$

The space $L^1(X,\mathcal{S},p)^*$ of all linear functionals on $L^1(X,\mathcal{S},p)$ is a normed space with the norm $\|\phi\|$ defined by

$$\|\phi\| = \sup_{\|f\|_1\le 1} |\phi(f)|$$

Recall (see e.g. Chapter 3 in [12]) that the space $L^1(X,\mathcal{S},p)^*$ is isomorphic to $L^\infty(X,\mathcal{S},p)$ (with the $\|\cdot\|_\infty$-norm); that is, there is an isometric isomorphism $h : L^1(X,\mathcal{S},p)^* \to L^\infty(X,\mathcal{S},p)$. The h-image of ϕ_q is the Radon–Nikodym derivative $\frac{dq}{dp}$ of q. Now, we have

$$\begin{aligned} d_3(q,r) &= \sup_{\|f\|_1\le 1} \left| \int f dq - \int f dr \right| = \sup_{\|f\|_1\le 1} |\phi_q(f) - \phi_r(f)| \\ &= \|\phi_q - \phi_r\| = \|h(\phi_q) - h(\phi_r)\|_\infty = \sup_{x\in X} \left| \frac{dq}{dp}(x) - \frac{dr}{dp}(x) \right| = d_4(q,r) \end{aligned}$$

To complete the proof it is enough to show that d_1 and d_3 are equivalent.

$$\begin{aligned} d_1(q,r) &= d_2(q,r) = \sup_{E\subseteq X} \left| \int \chi_E dq - \int \chi_E dr \right| \overset{\|\chi_E\|_1 \le 1}{\le} \sup_{\|f\|_1\le 1} \left| \int f dq - \int f dr \right| \\ &= \sup_{\|f\|_1\le 1} \left| \sum_{x\in X} f(x)(q(\{x\}) - r(\{x\})) \right| \le \sup_{\|f\|_1\le 1} \sum_{x\in X} |f(x)||(q(\{x\}) - r(\{x\}))| \\ &= \sup_{\|f\|_1\le 1} \left(\sum_{x\in X} |f(x)| \cdot \sum_{x\in X} |q(\{x\}) - r(\{x\})| \right) \le 2d_1(q,r) \cdot \sup_{\|f\|_1\le 1} \sum_{x\in X} |f(x)| \end{aligned}$$

Now $\|f\|_1 \le 1$ means $\sum_{x\in X} |f(x)|p(x) \le 1$ and thus there is a constant (depending only on p) such that $\sum_{x\in X} |f(x)| \le K$ holds for all $\|f\|_1 \le 1$. Therefore we obtained

$$d_1(q,r) \le d_3(q,r) \le 2Kd_1(q,r)$$

which completes the proof. ∎

Recall that for finite X, a sequence $(q_n) \subseteq M(\mathcal{S})$ of measures is said to *weak*-converge* (cf. [1][p. 335]) to $q \in M(\mathcal{S})$ if for all $f : X \to \mathbb{R}$ we have

$$\left|\sum_{x\in X} f(x)\big(q_n(\{x\}) - q(\{x\})\big)\right| \longrightarrow 0 \quad \text{as } n \to \infty$$

The topology of weak*-convergence is the same as the topology generated by d_1 (and hence by any of the d_i's). Indeed, suppose q_n weak*-converges to q. Choose $f = 1$. Then

$$\left|\sum_{x\in X} \big(q_n(\{x\}) - q(\{x\})\big)\right| = 2d_1(q_n, q) \longrightarrow 0 \quad \text{as } n \to \infty$$

Thus we see that it does not matter which of the metrics d_i we use when studying topological properties of $M(\mathcal{S})$. For convenience we let $(M(\mathcal{S}), d)$ to be a metric space with

$$d(q, r) = \max_{x\in X} |q(\{x\}) - r(\{x\})|$$

Later we will need the following Lemma, proof of which is left to the reader.

Lemma 4.3. For all $\varepsilon > 0$ there is $\delta > 0$ such that

$$\max_{x\in X} |q(\{x\}) - r(\{x\})| < \delta \quad \Longrightarrow \quad \max_{x\in X} \left|\frac{q(x)}{p(x)} - \frac{q'(x)}{p(x)}\right| < \varepsilon.$$

4.2 Proof of Proposition 3.2

Proof.

Recall that q is not Bayes learnable if and only if $\frac{dq}{dp}$ is injective (Proposition 2.1). Also recall that the Radon–Nikodym derivative $\frac{dq}{dp}$ is the function $x \mapsto \frac{q(\{x\})}{p(\{x\})}$. Let $BBS(p)$ denote the Bayes p-Blind Spot

$BBS(p)$ **is open:** Take any $q \in BBS(p)$. We shall prove that there is $\delta > 0$ such that for any $q' \in M(\mathcal{S})$ with $d(q, q') < \delta$ we have $q' \in BBS(p)$. Since $q \in BBS(p)$ the density function $\frac{dq}{dp}$ is not injective. It is enough to prove that for small enough δ, if $d(q, q') < \delta$, then $\frac{dq'}{dp}$ is not injective. Let

$$\varepsilon = \frac{1}{2} \min_{x\neq y} \left|\frac{q(\{x\})}{p(\{x\})} - \frac{q(\{y\})}{p(\{y\})}\right|$$

be the half of the minimal difference of different values of $\frac{dq}{dp}$. Injectivity of $\frac{dq}{dp}$ implies $\varepsilon > 0$. Using Lemma 4.3 there is $\delta > 0$ such that

$$d(q, q') < \delta \quad \text{implies} \quad \max_{x\in X} \left|\frac{q(x)}{p(x)} - \frac{q'(x)}{p(x)}\right| < \varepsilon$$

And this latter inequality ensure that $\frac{dq'}{dp}$ must be injective and thus $q' \in BBS(p)$.

$BBS(p)$ **is dense:** We need to verify that for all $q \in M(\mathcal{S})$ and $\delta > 0$ there is $q' \in BBS(p)$ such that $d(q, q') < \delta$. Let us fix q and δ and chose a function $\varepsilon : X \to (-\delta,\ \delta)$ such that $\sum_{x\in X} \varepsilon(x) = 0$ and if $q(\{x\}) = 0$, then $\varepsilon(x) > 0$. Define the measure q' on the singletons $x \in X$ by

$$q'(\{x\}) = q(\{x\}) + \varepsilon(x) \quad \text{for all } x \in X$$

Then q' is a probability measure as $\sum_{x \in X} \varepsilon(x) = 0$, and we obtain $d(q, q') < \delta$. It is straightforward to see that ε can be chosen in such a manner that

$$\frac{q(\{x\}) + \varepsilon(x)}{p(\{x\})} \neq \frac{q(\{y\}) + \varepsilon(y)}{p(\{y\})} \qquad \text{for all } x \neq y \in X$$

whence injectivity of $\frac{dq'}{dp}$ follows. Therefore $q' \in BBS(p)$ and the proof is complete.

∎

4.3 Proof of Proposition 3.4

Proof. Let $L \subseteq M$ be the set of Bayes learnable measures. We claim $\lambda^{n-1}(L) = 0$. $q \in M$ is Bayes learnable if and only if its Radon–Nikodym derivative $\frac{dq}{dp}$ is not injective, i.e.

$$L = \left\{ q \in M : \ \frac{q}{p}(x) \text{ is not injective} \right\}$$

Let $\mathcal{A} \subseteq \mathcal{P}(X)$ be a partition of X and call $\mathcal{A}$ a non-trivial partition if $\mathcal{A} \neq \big\{ \{x\} : \ x \in X \big\}$, i.e. there is at least one block $A \in \mathcal{A}$ that contains at least two elements. Write

$$L_{\mathcal{A}} = \left\{ q \in M : \ \frac{q(x)}{p(x)} \text{ is constant on every } A \in \mathcal{A} \right\} \ \subseteq \ M$$

Then we have

$$L = \bigcup \{ L_{\mathcal{A}} : \ \mathcal{A} \text{ is a non-trivial partition of } X \}$$

The proof proceeds as follows: we show that for each non-trivial partition $\mathcal{A}$ the dimension $\dim L_{\mathcal{A}}$ is at most $n-2$. Consequently $\lambda^{n-1}(L_{\mathcal{A}}) = 0$ and since there are only finitely many partitions of X, we obtain

$$\lambda^{n-1}(L) = \lambda^{n-1} \left(\bigcup L_{\mathcal{A}} \right) \leq \sum \lambda^{n-1} (L_{\mathcal{A}}) = 0$$

Take a non-trivial partition $\mathcal{A}$ and pick a block $A \in \mathcal{A}$ in this partition that contains at least two elements, say x_i and x_j. For any $q \in L_{\mathcal{A}}$ the function $\frac{dq}{dp}$ should be constant on A. The equation

$$\frac{X_i}{p(x_i)} = \frac{X_j}{p(x_j)} \qquad (X_i, X_j \in \mathbb{R} \text{ variables}) \tag{22}$$

defines a hyperplane H' and $L_{\mathcal{A}} \subseteq H'$. Clearly none of H' or H contains the other, therefore the intersection $H \cap H'$ has dimension at most $n-2$. Since $L_{\mathcal{A}} \subseteq H \cap H'$, we obtain the desired inequality $\dim L_{\mathcal{A}} \leq n-2$ and the proof is complete. ∎

References

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

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

[3] N.L. Carothers. *Real Analyis.* Cambridge University Press, Cambridge, 2000.

[4] John H. Conway and Richard Guy. *The book of numbers.* Copernicus – Springer, New York, 1996.

[5] Z. Gyenis and M. Rédei. Bayesian dynamical systems. 2016. in preparation.

[6] Z. Gyenis and M. Rédei. General properties of general Bayesian learning. *Erkenntnis*, 2016. submitted.

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

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

[9] J.C. Oxtoby. *Measure and Category*, volume 2 of *Graduate Texts in mathematics.* Springer-Verlag, New York Heidelberg Berlin, 2nd edition, 1980. First edition 1971.

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

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

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

[13] W. Rudin. *Functional Analysis.* International Series in Pure and Applied Mathematics. McGraw-Hill, New York, 2nd edition, 1991.

[14] Lynn Arthur Steen and Jr. J. Arthur Seebach. *Counterexamples in Topology.* Springer-Verlag, New York, 1978. Reprinted by Dover Publications, New York, 1995.

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