Updating Probability: Tracking Statistics as Criterion

Bas C. van Fraassen[1] and Joseph Y. Halpern[2]

Joe Halpe
Deleted:

1 Introduction

What is generally called Bayesian Conditionalization is a policy for updating probability assignments. It specifies as admissible input ("evidence") elements of the domain of the prior probability function, and allows as possible posteriors the conditionalizations on such elements ("propositions"). Under certain ideal conditions, this is the only coherent policy (Teller and Fine 1975, Teller 1976, Diaconis and Zabell 1982,van Fraassen 1999). When those conditions are not met, other policies might be appropriate. Putative examples include updating by Richard Jeffrey's generalized conditionalization ("Jeffrey Conditionalization") or Edwin T. Jaynes' rule to maximize relative entropy ("MAXENT"). This raises the question of what general constraints any such policy should satisfy. We will propose an answer, guided initially by some intuitive considerations that also motivate the Bayesian policy.

It is no coincidence that the probability calculus is typically introduced with examples of urns full of red and black, or marble and wooden, balls. The relations between the statistical proportions of finite sets, with their intersections and unions, introduce the basic principles that are extrapolated to present probability theory, and that extrapolation is motivated by the insistence that a probability assignment should in principle be able to track, in some appropriate sense, the relevant statistics. It is well known that the extrapolation goes far beyond the theory of finite proportions, and beyond the theory of limits of relative frequencies, but there are also far-reaching theorems to show that the relationship of these to probability theory remains appropriately close.[3]

That this possibility of tracking the relevant statistics should be preserved under updating also provides a motivation for the Bayesian policy. If the proportions are known, and we are given the information that a ball drawn from the urn is marble, then the probability that it is red should be updated from the proportion of red balls in the urn to the proportion of red balls among the marble ones in the urn. And if the proportions are not known, but the prior probabilities were assigned on the basis of opinion or evidence, then the updating should take the same form, in order to guarantee the following:

> If the prior probability assignment had the possibility that it was tracking the relevant statistics for draws of balls from the urn, then that possibility was not lost for updating to draws of balls from among the marble balls in the urn.

2 Alternative updating policies

Alternatives to the Bayesian policy have been discussed for two main reasons. The first is that the input that triggers a change in probabilities may not be of the sort this policy takes into account. Both the above examples, of Jeffrey Conditionalization and Jaynes' MAXENT, were introduced in this way. In his 'Probable Knowledge' Richard Jeffrey proposed to allow for input beyond the propositional, with the following example:

> In examining a piece of cloth by candlelight one might come to attribute probabilities 0.6 and 0.4 to the propositions *G* that the cloth is green and *B* that it is blue, without there being any proposition *E* for which the direct effect of the observation is anything near changing the observer's degree of belief in *E* to l. (Jeffrey 1968: 172)

The proper transformation of probability he offered, now known as Jeffrey Conditionalization, redistributes the probability over a partition whose cells are the affected alternatives:

$$P'(A) = \Sigma_i \, q_i P(A|C_i),$$

with $\{C_i : i = 1, 2, \ldots, n\}$ the relevant partition, such as *green, blue, ..., red*, and $\{q_i : i = 1, 2, \ldots, n\}$ the the weights of the new posterior probabilities of those alternatives.

Edwin T. Jaynes introduced his maximum entropy updating rule in (Jaynes 1957); a typical motivating example takes the following sort of input:

> We have a table which we cover with black cloth, and some dice, but … they are black dice with white spots. A die is tossed onto the black table. Above there is a camera [which] will record only the white spots. Now we don't change the film in between, so we end up with a multiple exposure; uniform blackening of the film

> after we have done this a few thousand times. From the known density of the film and the number of tosses, we can infer the average number of spots which were on top, but not the frequencies with which various faces came up. Suppose that the average number of spots turned out to be 4.5 instead of the 3.5 that we might expect from an honest die. Given only this information … what estimates should [we] make of the frequencies with which *n* spots came up? (Jaynes 2003: 343).

Transferring this problem into one of probability updating, the prior probability assignment, assuming a fair die, included an expectation value of 3.5 for the number of dots, and the input is an new expectation value of 4.5. There is no proposition in the domain of the prior on which to conditionalize; this input is of a different sort. What should be the posterior probability assignment? Jaynes' rule gives a precise answer that implies, for example, that the outcome with 1 spot has posterior probability 0.05 and the outcome with 6 spots has posterior probability 0.35.

The second reason that has entered the discussion of alternative policies is that there may be conditions seen as triggering a more radical change than can be accommodated by conditionalization. It may be a case, as it is often put, to "throw away the prior", but even so the change would not be a choice at random; the agent could be guided by prior experience and theoretical presuppositions that appear phenomenologically only as intuition and instinct, limiting the acceptable choices.

What we will investigate is a way to place requirements on any policy for updating that remains well motivated by the intuitive considerations offered above in terms of the possibility to track relevant statistics and of preserving that possibility. As Jaynes' example illustrates, such motivation remains salient when departures from the Bayesian policy are envisaged.

3 Modeling the situation for normal updating

A policy for updating probabilities needs to start with a description of the sort of situation to which it applies. Such a situation will be one where the subject or agent has at each time a state of opinion represented by a probability function on a space (representing a range of possibilities or possible circumstances). Secondly, it must be specified how, in this sort of situation, there can be input of various sorts. Thirdly, the policy must offer a prescription of what that prior probability function is allowed to be changed into, in response to a given input, to form a posterior probability function.

The Bayesian policy, applicable here, starts with a prior probability, takes the inputs to be elements of a finite partition of the possibilities, and, given an element of the partition, updates the prior to a posteriur by conditionalization. We will not assume that

in general a policy must prescribe a unique posterior for each possible input, nor that that the posteriors correspond to cells in a partition.

To accommodate not just the Bayesian policy but also the cases presented by, for example, Richard Jeffrey and Edwin Jaynes, we must not concentrate on the special case in which the input is an 'evidential' proposition in the domain of the prior probability assignment. Indeed, Jeffrey's example introduces an agent who has no way of expressing what triggers the change in probability assignment, hence no input to which a conscious policy could apply. Nevertheless, it can be specified that this change, and the agent's response when managing his overall opinion, takes a very special form. We note that in this case too, given a prior, the input, whether explicit or unformulated, places a constraint on the posterior. An input, whatever it may be, acts as a function that takes any given prior into a set of possible (admissible, acceptable) posteriors. As the most general form of input we must therefore count any constraint that the agent may accept as limiting the candidates for the posterior probability function. Whatever types of input there can be can therefore be represented by functions that take any element of the space (a prior) into a subset of that space (its set of possible posteriors).

A little more formally: a policy for updating (whether or not in response to inputs representable in terms of the elements of the space of possibilities) specifies for each probability function $\boldsymbol{p}$ (the 'prior') a set $\boldsymbol{R}$ of possible 'posteriors' (of cardinality greater than 1), or equivalently, a constraint on the functions into which $\boldsymbol{p}$ can change by updating. We place only one condition on how these responses are formed. Since we view the process of going from $\boldsymbol{p}$ to $\boldsymbol{R}$ as normal updating, not revolutionary change, we assume ,: updating does not raise any probability from zero. For succinctness, a model (of a doxastic situation, in the present context) is a triple $M = \langle \boldsymbol{p}, \boldsymbol{S}, \boldsymbol{R} \rangle$, where $\boldsymbol{p}$ is a probability measure on $\boldsymbol{S}$ and $\boldsymbol{R}$ is a set of probability measures on $\boldsymbol{S}$ that assign 0 wherever $\boldsymbol{p}$ assigns 0. As stipulated above, the number of possible posteriors, the members of $\boldsymbol{R}$, is greater than 1.[4]

4 Tracking: a criterion for updating policies

Following upon the intuitive motivation presented above we propose a formal criterion to be met by updating policies. The motivating considerations included two main points: opinion represented in terms of a probability assignment should at least possibly track the relevant statistics, and updating the probability assignment on new input should preserve that possibility. We added that under certain ideal conditions, Bayesian Conditionalization is precisely the policy that satisfies these requirements.

What the ideal conditions are, and how more practical conditions could be related to the ideal case, is illustrated in the following (partly fictional) example.

In 1994 the distribution, by party and gender, in the United States Senate was as follows:

	MEN	WOMEN	row total
REPUBLICAN	42	2	44
DEMOCRAT	51	5	56
column total	93	7	100

Suppose a fully informed ideal agent A* makes all the relevant distinctions and actually has the information about each individual Senator, about their gender, age, party affiliation, and more, like their state of health, etc. Assume that his probabilities match these statistics, for example,

P(x is a woman | x is a Republican Senator) = 2/44.

A certain agent A, who is not a fully informed ideal agent, let us assume, has no information about state of health, age, or even gender; only about party affiliation in the Senate at that time. Within this very limited space, his probabilities match the statistics too, but lack the male/female distinction: he just has, for example,

P'(x is a Republican | x is a Senator) = 44/100.

Thus the probabilities that A assigns are the *marginal probabilities* assigned by A*, matching the marginal distribution in the above table with the division by gender removed.

Now the fully informed ideal agent A* gets the new information that there are no more women Senators, they were all removed from office for such reasons as fraud, health, or other personal circumstances. Being an ideally well-informed agent, one that not only has all the relevant statistical information for his subject matter, but is also making all the relevant distinctions, the Bayesian policy is not only applicable but the uniquely right policy to follow. Thus A* conditionalizes on this information, reclassifying those women as no longer in the Senate. His posterior probabilities match:

	MEN	WOMEN	row total
REPUBLICAN	42	0	42
DEMOCRAT	51	0	51
column total	93	0	93

Suppose now that the non-fully informed ideal agent ***A***, who does not make all the same distinctions as *A**, and lacks either information or opinion about at least some of the aspects to which ***A**** is privy, nevertheless has a prior and posterior opinion that remain

correctly represented by marginals of the prior and posterior of the fully informed ideal agent. This means that we see him updating (through whatever impulse or choice) to the corresponding marginal distribution: Republican 42, Democrat 51. So now both ***A**** and ***A*** have, for example, the probability 42/93 for a Senator being a Republican.

Watching this subject A who is not ideally well-informed, we see that his change of opinion *was not one done by conditionalization*. All we see is a redistribution over the same alternatives. From the outside, it was just an abrupt change in the probabilities for party affiliation in the Senate. However, since his posterior probability is the marginal of an updating by conditionalization by the fully informed ideal subject, the virtue of matching the actual statistics was preserved.

So the criterion for updating policies that we propose will be that however the agent assigns his probabilities and updates them, there exists in principle an ideal take on his situation, and his own opinion is in a relevant sense 'part of' the evolving probability assignment of an ideally well-informed (hence Bayesian) agent. Echoing the intuitive discussion, we shall call this criterion **Tracking.** Under what conditions is this criterion satisfied?

The question is reminiscent of discussions of hidden variable interpretations of quantum mechanics, in which unfamiliar sorts of stories, involving strange or unpredictable changes, are shown to be compatible with 'larger' stories that follow a familiar pattern. Specifically, conditions are investigated in which an assignment of probabilities of outcomes can have an 'underlying' classical probability model. Taking our cue from the above example, we take it that for a model to satisfy the **Tracking** criterion, its probability functions must be the *marginals* of probability functions in a *larger associated model* on which much stricter constraints may be imposed.[5]

In what follows we shall first show that **Tracking** is satisfied if and only if the prior probability assignment is a convex combination of the possible posterior assignments. Then secondly we will formulate a distinct criterion, **Spanning**, to govern evolving expectation values, and show that it is equivalent to **Tracking**. With those results in hand we will then show the diversity of updating policies that can satisfy those criteria, but also that there are policies which violate them.

5 Tracking: precise formulation and relation to convexity

In our example, we were looking at a single possible updating. For the fully informed ideal subject there were many possible posteriors, each of them a conditionalization on some proposition that changes one of the four numbers for his four 'cells'. And the intuitive Tracking criterion requires that the non-fully informed ideal subject's prior and possible posteriors should be the marginals of *some* such (imaginary) fully informed ideal subject's prior and possible posteriors.

We can concentrate on a representative simple form for such a case. Throughout we will take the set of possible posteriors to be finite. But as will be clear in retrospect, the updating policy could allow for all the convex combinations of a given finite set as the set of possible posteriors.

Definition. A model $M^* = \langle \boldsymbol{p}^*, \boldsymbol{S}^*, \boldsymbol{R}^* \rangle$ is an *associate* of model $M = \langle \boldsymbol{p}, \boldsymbol{S}, \boldsymbol{R} \rangle$ iff there is an integer n such that:

i. $\boldsymbol{R} = \{ \boldsymbol{p}_1, \ \ldots, \boldsymbol{p}_n \}$;
ii. $\boldsymbol{S}^* = \boldsymbol{S} = \{1, \ldots, n\}$;
iii. $\boldsymbol{p}$ is the marginal of $\boldsymbol{p}^*$ on $\boldsymbol{S}$, i. e., $\boldsymbol{p}(E) = \boldsymbol{p}^*(E \otimes \{1, \ldots, n\})$ for each measurable subset E of $\boldsymbol{S}$;
iv. $\boldsymbol{R}^* = \{ \boldsymbol{p}_1^*, \ldots, \boldsymbol{p}_n^* \}$ defined by:
    a. $\boldsymbol{p}_j^*(\boldsymbol{S} \otimes \{j\}) = 1$ [for $j = 1, \ldots, n$];
    b. for each $\boldsymbol{p}'$ in $\boldsymbol{R}$ there is a single member $\boldsymbol{p}'^*$ of $\boldsymbol{R}^*$ such that $\boldsymbol{p}'$ is the marginal of $\boldsymbol{p}'^*$, i.e. $\boldsymbol{p}'(E) = \boldsymbol{p}'^*(E \otimes \{1, \ldots, n\})$.

At this point, we have assumed nothing about the relation between $\boldsymbol{p}^*$ and the family $\{\boldsymbol{p}_1^*, \ldots, \boldsymbol{p}_n^*\}$ beyond the fact that the latter are absolutely continuous with respect to the former. This implies that $p^*(\boldsymbol{S} \otimes \{j\}) > 0$ for $j = 1, \ldots, n$.

The set $\{1, \ldots, n\}$ could represent a partition of a larger space (the 'hidden parameters'), but what these integers represent is not relevant to the argument; in our proofs, they function simply as an index set. We can think of the members of $\boldsymbol{R}$ as also indexed by these integers, in a corresponding fashion. Let $\boldsymbol{p}_j$ be the member of $\boldsymbol{R}$ such that $\boldsymbol{p}_j^*$ is the probability function in $\boldsymbol{R}^*$ indicated in clause iii(b). Then $\boldsymbol{p}_j(E) = \boldsymbol{p}_j^*(E \otimes \{j\})$ for each subset E of S.

The only constraint on the measure $\boldsymbol{p}^*$ in the definition above is given by clause (iii), so it is clear that each model M will have many associate models. This allows the introduction of the guiding constraint **[Tracking]** in these terms:

> Model $M = \langle \boldsymbol{p}, S, R \rangle$ satisfies [**Tracking**] if and only if M has an associated model $M^* = \langle \boldsymbol{p}^*, \boldsymbol{S}^*, \boldsymbol{R}^* \rangle$ in which $\boldsymbol{p}^*$ is a probability function on $\boldsymbol{S}^*$ such that the members of $\boldsymbol{R}^*$ are conditionalizations of $\boldsymbol{p}^*$, specifically, $\boldsymbol{p}_j^*(\mathrm{E} \otimes \{j\}) = \boldsymbol{p}^*(\mathrm{E} \mid \boldsymbol{S} \otimes \{j\})$ for $j = 1, \ldots, n$).

Let us spell out precisely what this involves.[6]

Suppose that M, as described, satisfies [**Tracking**], with M^* the associate model, and that $\boldsymbol{R}$ is countable. Then $\boldsymbol{p}^*$ is a convex combination of the members of $\boldsymbol{R}^*$, the weights being the probabilities $\boldsymbol{p}^*(\boldsymbol{S} \otimes \{j\})$ for $j = 1, \ldots, n$, each of these being positive.

Since p^* gives positive probability to each set $\{n\}$, and N has at least two members, it follows that these are *strict* convex combinations, that is, the values of $\boldsymbol{p}$ for elements of $\boldsymbol{S}^*$ are neither infimum nor supremum of the corresponding values assigned by members of $\boldsymbol{R}^*$, unless those values are all the same.[7]

The marginal $\boldsymbol{p}$ of $\boldsymbol{p}^*$ on $\boldsymbol{S}$ is:

$$\begin{aligned}
\boldsymbol{p}(E) &= \boldsymbol{p}^*(E \otimes \{1, \ldots, n\}) \\
&= \Sigma_j \boldsymbol{p}^*(E \otimes \{j\}) \\
&= \Sigma_j \boldsymbol{p}^*(E \otimes \{j\} \mid \boldsymbol{S} \times \{j\})\, \boldsymbol{p}^*(\boldsymbol{S} \otimes \{j\}) \\
&= \Sigma_j \boldsymbol{p}_j^*(\mathrm{E} \otimes \{j\})\, \boldsymbol{p}^*(\boldsymbol{S} \otimes \{j\}) \\
&= \Sigma_j \boldsymbol{p}_j(\mathrm{E})\, \boldsymbol{p}^*(\boldsymbol{S} \otimes \{j\}).
\end{aligned}$$

So $\boldsymbol{p}$ is that same convex combination of the members of $\boldsymbol{R}$. Therefore:

Theorem 1. If $M = \langle \boldsymbol{p}, \boldsymbol{S}, \boldsymbol{R} \rangle$, with $\boldsymbol{R}$ countable, satisfies [**Tracking**], then $\boldsymbol{p}$ is a strict convex combination of the members of $\boldsymbol{R}$.

The converse to Theorem 1 also holds. Suppose that $\boldsymbol{p} = \Sigma_j\, c_j \boldsymbol{p}_j$, with the weights c_j non-negative and summing to 1. We can then construct an associate model M^* as above, filling in the single 'blank' left in the definition by setting $\boldsymbol{p}^* = \Sigma_j\, c_j \boldsymbol{p}_j^*$ taking $\boldsymbol{p}^*(\boldsymbol{S} \otimes \{j\}) = c_j$, for $j = 1, \ldots, n$. It is easy to see then that $\boldsymbol{p}$ is the marginal of $\boldsymbol{p}^*$ on $\boldsymbol{S}$, since that is how the probabilities $\boldsymbol{p}^*(\boldsymbol{S} \otimes \{j\})$ were chosen. Thus:

Theorem 2. If in $M = \langle \boldsymbol{p}, \boldsymbol{S}, \boldsymbol{R} \rangle$, with $\boldsymbol{R}$ countable, the prior $\boldsymbol{p}$ is a strict convex combination of the members of $\boldsymbol{R}$, then M satisfies [**Tracking**].

6 The Spanning Criterion

It would be preferable to have a formulation of the criterion that can be applied directly to the relation between the prior and the possible posteriors (allowed by the policy in question, in the situation in question), without recourse to a study of other models.

For this we offer the following criterion, which has some history in the literature:

> Model $M = \langle \boldsymbol{p}, \boldsymbol{S}, \boldsymbol{R} \rangle$ satisfies [**Spanning**] if and only, for each random variable (r.v.) g, $E_{\boldsymbol{p}}[g]$, the expected value of g with respect to $\boldsymbol{p}$, lies strictly inside the interval spanned by $E_{\boldsymbol{p}'}[g]$ for $\boldsymbol{p}' \in \boldsymbol{R}$.

Again, 'strictly inside' an interval means that it is neither infimum nor supremum unless the interval is a single point.[8]

We note that [**Spanning**] is the same as the *general reflection principle* (van Fraassen 1995, 1999), except that, in the present formulation, there is no reference to prior probabilities about what the future probabilities will be like. With suitable assumptions, the requirement to satisfy [**Spanning**] implies the original Reflection Principle (van Fraassen 1995: 18-19) and under special conditions coincides with the Bayesian updating policy (van Fraassen 1999: 96).[9]

Theorem 3. If $M = \langle \boldsymbol{p}, \boldsymbol{S}, \boldsymbol{R} \rangle$ satisfies [**Tracking**], and both $\boldsymbol{S}$ and $\boldsymbol{R}$ are countable, then M satisfies [**Spanning**].

If M satisfies [**Tracking**] then, by Theorem 1, $\boldsymbol{p}$ is a convex combination $\Sigma_{j \in N}\, c_j \boldsymbol{p_j}$ of the members of $\boldsymbol{R}$, with all coefficients positive. Suppose that g is an r.v.; since S is countable, the expected value of g with respect to $\boldsymbol{p}$ is

$$\begin{aligned} E_{\boldsymbol{p}}[g] &= \Sigma_{x \in \boldsymbol{S}}\, \boldsymbol{p}(x) g(x) \\ &= \Sigma_{\, x \in \boldsymbol{S},\, \Sigma_j\, c_j \boldsymbol{p_j}}(x) g(x) \\ &= \Sigma_j\, c_j\, (\Sigma_{x \in \boldsymbol{S}}\, \boldsymbol{p_j}(x) g(x)) \\ &= \Sigma_j\, c_j E_{\boldsymbol{pj}}[g]. \end{aligned}$$

So $E_{\boldsymbol{p}}[g]$ is a convex combination of the expected values of g with respect to the members of $\boldsymbol{R}$, and since all the coefficients are positive, lies strictly inside the interval spanned by the latter.

Theorem 4. If $M = \langle \boldsymbol{p}, \boldsymbol{S}, \boldsymbol{R} \rangle$, $\boldsymbol{S}$ and $\boldsymbol{R}$ are finite, and $|\boldsymbol{R}| > 1$, satisfies [**Spanning**], then M satisfies [**Tracking**].

In view of Theorem 2, it suffices here to prove the following:[10]

Lemma. If M satisfies [**Spanning**], where $\boldsymbol{S}$ and $\boldsymbol{R}$ are finite, and $|\boldsymbol{R}| > 1$, then $\boldsymbol{p}$ is a strict convex combination of the members of $\boldsymbol{R}$.[11]

The random variables defined on $\boldsymbol{S}$ are the functions that map S into the real numbers. These form a vector space (isomorphic to the familiar R^n, the vector space of n-tuples of real numbers)
with addition and scalar multiplication defined point-wise:

$$cg(x) = c(g(x)); \qquad (g + f)(x) = g(x) + f(x).$$

The probability measures form a convex subset of this vector space, defined by

$\boldsymbol{q}(x) \geq 0; \quad \Sigma_{x \in S} \boldsymbol{q}(x) = 1.$

The fact that the set is convex follows from the observation that if $\boldsymbol{p}$ and $\boldsymbol{q}$ are probability measures on S, then so is $c\boldsymbol{p} + (1\text{-}c)\boldsymbol{q}$, for c in [0,1].

The expected value of g with respect to a probability measure $\boldsymbol{q}$ is, in vector space terms, just the scalar product:

$(\mathrm{g}, \boldsymbol{q}) = \Sigma_{x \in S} \boldsymbol{q}(x)\mathrm{g}(x).$

If $\boldsymbol{R}$ is finite, the convex hull [$\boldsymbol{R}$] of $\boldsymbol{R}$ (i.e., the set of all probability measures that are convex combinations of the elements of $\boldsymbol{R}$) is called a *polytope*. The extreme points are those that are not convex combinations of other members; a polytope has extreme points. Since $\boldsymbol{R}$ has more than one member, it has at least two extreme points and also non-extreme points. (The simplest case is the one where $\boldsymbol{R}$ has just two members, which are the extreme points of [$\boldsymbol{R}$].)

We need to show that unless $\boldsymbol{p}$ is in [$\boldsymbol{R}$] but not an extreme point of [$\boldsymbol{R}$], then [**Spanning**] is violated. The latter means that there is some r.v. g such that the scalar product $(g, \boldsymbol{p})$ is not strictly in the interval spanned by the set $\{(\mathrm{g}, \boldsymbol{p'}): \boldsymbol{p'} \in [\boldsymbol{R}]\}$.

Two vectors are g, h are *orthogonal* iff $(g, h) = 0$. In the familiar three-dimensional vector space, the two-dimensional subspaces are the planes, and a plane is the set of vectors orthogonal to a given vector. So a plane is defined by an equation of the form $(g, \boldsymbol{q}) = 0$ for a fixed vector g; that is, $\boldsymbol{q}$ is in the plane iff $\Sigma_{x \in S} g(x)\boldsymbol{q}(x) = 0$. In our context, where $\boldsymbol{q}$ is a probability function, this means that the expected value of g with respect to $\boldsymbol{q}$ is 0.

In general, the maximal proper subspaces of a vector space are called the hyperplanes, and again, any such subspace is the ortho-complement of a single vector, defined in the same way. If H is the hyperplane $\{h: (g, h) = 0\}$, then it divides the space into two half-spaces, overlapping only in H itself: namely $\{h: (g, h) \leq 0\}$ and $\{h: (g, h) \geq 0\}$. A polytope is the intersection of a finite set of half-spaces, and equivalently, the convex hull of a finite set of vectors. Its extreme points are those that are not convex combinations of other members. A polytope has a finite set of extreme points, all of which are in it, and it is the convex hull of its set of extreme points.

Since $\boldsymbol{R}$ is finite, [$\boldsymbol{R}$] is a polytope.[12] Let [$\boldsymbol{R}$] thus be the intersection of a finite set of *half-spaces* $H_1, \ldots, H_\mathrm{m}$ defined by the inequalities

$(h_1, x) \leq 0, \ldots, (h_\mathrm{m}, x) \leq 0.$

The corresponding *supporting hyperplanes* $T_1, \ldots, T_\mathrm{m}$, which contain the faces of [$\boldsymbol{R}$], are defined by the equalities

$(h_1 , x) = 0, ..., (h_m , x) = 0.$

Now suppose that $\boldsymbol{p}$ is not a strict convex combination of members of $\boldsymbol{R}$. Then $\boldsymbol{p}$ is either outside [$\boldsymbol{R}$] or else is an extreme point of [$\boldsymbol{R}$]. If $\boldsymbol{p}$ is outside [$\boldsymbol{R}$] then there is a half-space H_j to which $\boldsymbol{p}$ does not belong. Hence $(h_j , \boldsymbol{p})$ is positive, and thus greater than (h_j , x) for any member x of [$\boldsymbol{R}$].

If $\boldsymbol{p}$ is an extreme point of [$\boldsymbol{R}$], then it is the unique intersection of a sub-family of the hyperplanes above:

$$\{\boldsymbol{p}\} = T_{i1} \cap ... \cap T_{ik}$$

(cf. Gruber 2007: 246-247). So the random variable $\boldsymbol{g} = h_{i1} + ... + h_{ik}$ takes value 0 on $\boldsymbol{p}$. But if x is any other point in [$\boldsymbol{R}$], it will lie in all the corresponding half-spaces, but not on all of these hyperplanes, hence $g(x) < 0$. Hence the expected value of $\boldsymbol{g}$ with respect to $\boldsymbol{p}$ is not *strictly* in the interval spanned by the expected values of $\boldsymbol{g}$ with respect to members of $\boldsymbol{R}$. Thus, in that case, [**Spanning**] is violated.

7 Non-Bayesian policies that satisfy the Spanning/Tracking criterion

The close connection between [**Tracking**] and conditionalization may give the impression that the results proved here impose Bayesian conditionalization as the sole admissible policy. That is not so, for a number of reasons.

The orthodox Bayesian policy is this:

- accept as admissible input only propositions;
- as response to such an input the only admissible change is conditioning the prior on the proposition in question.

So an updating policy can depart from the Bayesian in one or more of three ways:

1. accept as admissible a wider variety of inputs (e.g. expected values);
2. an admissible response to such an input can be a change in the prior that is not the result of conditioning;
3. an admissible response to such an input may be non-unique, that is, the posterior may not be uniquely determined by the prior + input.

For example, therefore, a policy could be non-Bayesian by differing in the third way, even if the posterior is formed from the prior by conditionalizing. That could be so if the updating involved a choice, or instead of a free choice, involved some factor in addition to prior and explicit 'evidential' input that helps to determine the posterior – a 'hidden variable', of which the agent might or might not be aware.

Second, we have included no initial assumption here about how the possible posterior opinions relate to the possible inputs. It is not assumed that for each input there is only one admissible posterior opinion, nor that they are conditionalizations of the prior.

Third, the embedding in a larger probability space (associate model) is in no way unique. The 'hidden variables' are underdetermined. Nor is there any reason to think that there the same larger probability space would be pertinent for transitions over subsequent time intervals $(t, t + a)$, $(t + a, t + a + b)$, $(t, t + a + b)$.

Fourth, the possible posteriors in a model that satisfies [**Tracking**] are in general *not* conditionalizations of the prior. That was quite clear in our 'Senate' example above. For the marginals of conditionalizations of a probability function are only in special cases conditionalizations of a marginal. A special case is the one in which the 'hidden variables' over which an average is taken are actually independent of the 'surface' variables.

Finally, the criteria can with minor modifications in phrasing be applied to policies governing interval-valued rather than sharp probabilities, where it is still even far from clear what the proper analogue to Bayesian conditionalization must be.

It is helpful to look at some simple examples of how [**Spanning/Tracking**] can be satisfied by non-Bayesian updating policies. Our first illustration already provided a good example: a policy that consists in adopting as probabilities certain marginals of the probabilities assigned by Bayesian conditionalization in a larger space. That sort of example can illustrate the various differences. But we add here two more examples of updating policies that involve some leeway, and are clearly not the Bayesian conditionalizing policy, but satisfy [**Spanning/Tracking**].

To begin, consider Jeffrey Conditionalization. Recall Jeffrey's example, cited above, of the agent who examines a piece of cloth by candlelight, and acts on an input which does not take the form of a proposition on which he could conditionalize his probability function. In response to this input, the agent redistributes the probability over a partition whose cells are the affected alternatives:

$$P'(A) = \Sigma_i \, q_i P(A|C_i),$$

with $\{C_i : i = 1, 2, \ldots, n\}$ the relevant partition, such as *green, blue, ..., red*, and the weights $\{q_i : i = 1, 2, \ldots, n\}$ the new posterior probabilities of those alternatives. If the agent's policy in this case dictates no constraints on those new weights then [**Spanning**] is satisfied, since the prior P is itself among the possible posteriors. Thus Jeffrey's proposal, by itself, is for a policy that satisfies our criteria.

We can imagine that certain other factors in the policy do place constraints on the new weights, in which case [**Spanning**] could be violated. When observing cloth by candlelight this would happen, for example, if the agent could only raise his probability that the cloth is green, and not lower it.

We can imagine a more complex situation in which the transformation requires not only a selection of new weights on a given partition, but also a choice of relevant partition. This would increase the set of possible posteriors accordingly. Jeffrey's example would become something like this:

> In examining a piece of cloth by candlelight one might come to attribute probabilities 0.6 and 0.4 to the propositions *G* that the cloth is green and *B* that it is blue, *or alternatively* come to attribute probabilities 0.7 and 0.3 to the propositions *C* that the cloth is cotton and *L* that it is linen, without there being any proposition *E* for which the direct effect of the observation is anything near changing the observer's degree of belief in *E* to l.

The criterion that [**Spanning**] be satisfied applies to the policy that allows these changes under those epistemic circumstances, rather than to the specific or actual change of posterior weights in either partition. Thusj **[Spanning]** is satisfied here.

As above, violation is possible if that policy has some further features that prevent a change in probability in one direction, either upward or downward, while allowing a change in the other. While it is on the face of it hard to see how a policy could appear rational while doing so, we shall see below that a well-known updating policy does exactly that.

For the next two examples, staying rather close to the Bayesian format, we consider the case where the agent is going to make an observation and knows that the event to be observed is a member of the partition $\{E_j : j \in J\}$. The first policy dictates conditionalizing the prior $\boldsymbol{p}$ on event E_k if the agent witnesses that event. But if the agent is not sure whether the event witnessed was, say, E_k or E_m, then the policy dictates that he Jeffrey conditionalizes, that is, adopt as posterior a convex combination of $\boldsymbol{p} \mid E_k$ and $\boldsymbol{p} \mid E_m$. But the policy does not dictate the weights in that convex combination, which can be chosen spontaneously from some given finite set. Hence *R* consists of certain convex combinations of the result of conditioning *p* on members of the partition, so [**Spanning**] is satisfied. This example already involves all three of the departures from the orthodox Bayesian mold: the input is not a proposition but something involving two propositions (with the agent's attitude not belief but ambivalence) and a choice of weights, so that the posterior is not uniquely determined by the input and prior.

A still different policy links posteriors to both the event witnessed and the agent's frame of mind, which is not determined by either the prior or the witnessed events. Suppose

that $\boldsymbol{p}_j$ is the probability that the agent uses when in frame of mind j, for j in some index set J. That is, the probability he assigns to events depends on his mood or on the context, as characterized by his frame of mind. As long as he stays in that frame of mind j, he updates by conditionalizing, but if his frame of mind changes from frame j to frame k, he updates by conditioning on $\boldsymbol{p}_k$. The agent might not even be aware that he assigns different probabilities in different frames of mind.

Does this last policy satisfy [**Spanning**]? It does. Suppose that the prior is $\boldsymbol{p}_1$ and the agent is about to begin an experiment with outcomes that yield the mutually exclusive propositions E_m for $m \in M$. It is possible that the agent stays in frame of mind 1, but it is also possible that he will shift into some frame j in J before conditioning. The set of possible posteriors is $\{\boldsymbol{p}_1(\ |E_m)\colon m \in M\}$, of which $\boldsymbol{p}_1$ is a convex combination.

8 Policies that violate the Tracking/Spanning criterion

We turn now to the limits set by the Tracking criterion.

A simple example will show that a violation of our criteria may be quite salient in an ordinary, easily imaginable situation. Imagine a doctor who announces to a patient that the probability that he has a certain virus is x. He prescribes a blood test that he says has, no adverse effects of any sort. Then he announces that if the test outcome is positive he will conclude that probability is at least twice as high, but if the outcome is negative they won't know any more than before, his current opinion will not change, and they will need more tests. (Various acquaintances of ours readily imagined that their doctors might announce something of this form.)

Only a little reflection shows that the doctor's probabilities, as they can evolve over this interval of time, cannot possibly be in accordance with the actual statistics throughout. If his prior probability, before the test is given, matches the statistics, so that x% of the population has the virus, how could the proportions in the three sub-populations (test positive, test negative, not tested) be $2x$%, x%, and x%? It is clear that to satisfy the **Spanning** criterion, the prior value *needs to fall strictly inside* the interval spanned by the posterior values.

Orthodox Bayesians would have the same complaint against the doctor, for this test situation, with its 'definite' propositional outcomes, is their paradigm example, and he is supposed to conditionalize (and to know beforehand that his possible posteriors will be conditionalizations on the outcomes). Thus the orthodox Bayesian policy of conditionalizing on new 'evidence' propositions satisfies [**Spanning**]. Currently, Objective Bayesians recognize a wider array of possible input forms, and follow the policy of maximizing relative entropy (MAXENT) originally proposed by Jaynes (cf. Williamson 2011). The debate between orthodox and objective Bayesians goes back to the 1970s. Specifically, Myron Tribus and Hector Motroni (1972) and Kenneth

Friedman (1973) debated an example in statistical physics where MAXENT violates [**Spanning]**.

Friedman insists on the criterion, though without naming it as a specific criterion, or elaboration, viewing it as corollary to the Bayesian policy:

> According to a Bayesian account that a prior probability cannot be revised upward … but will with probability p [> 0] be revised downward, implies that … the prior probability must be too high. (Friedman 1973: 266)

The currently more familiar, and simpler, 'Judy Benjamin' problem illustrates the same difficulty [van Fraassen 1981; Grove and Halpern 1997; Hartmann and Rad 2012].

In such examples, where the input is a new value for a conditional probability of a given event A on assumption B, there will be an element C disjoint from $A \cup B$ whose probability MAXENT will definitely raise or keep equal, regardless of what that new input value is (cf. van Fraassen 1981; the point is generalized by Seidenfeld 1987, Corollary 1, p. 283). If those probabilities are the relevant possible posteriors then [**Spanning**] is violated; therefore, by Theorem 3, [**Tracking**] is violated as well.

Many discussions have shown how this difficulty can disappear if the situation is described differently, either by adding information to the input, or constraining the range of possible inputs not ruled out by the prior, or constraining the range of possible posteriors allowed by the policy. Each of these alters the situation; Jaynes' prescription was specifically for the case in which there is nothing else to go on. Even today MAXENT remains controversial (see e.g. Grünwald and Halpern 2003). On the one hand, there is an explicit defense of MAXENT's violation of [**Spanning**] by Jon Williamson [2011: 68, 72, 80-81]; on the other hand, Brian Skyrms concluded on the basis of his results about the relation between MAXENT and conditionalization that 'MAXENT escapes dynamic incoherence by a hair's breadth' (Skyrms 2013: 82; see also his 1985, 1987).

APPENDIX. Generalizing to the countably infinite case

In the proof of Theorem 4, the underlying space S was assumed finite, so that the set of random variables formed a finite-,dimensional vector space. In the case where S is countable it becomes less clear how the criterion of possibly tracking real statistics is to be understood. Presumably, 'tracking the real statistics' would need to be cashed out in terms of matching limits of relative frequencies in countable sequences of samplings or events.

However, the proof of Theorem 4 can be generalized to include the case of S countable. In that case the vector space whose elements are the random variables defined on S, including the probability functions, must be chosen such that the scalar products (hence, the expected values) are well defined.

Let $S = \{x_j : j = 1, 2, 3, ...\}$. We restrict to *well-behaved* random variables, that is, the functions g such that $\Sigma_j|g(x_j)|^2$ is finite ('square integrable functions'). Then the random variables form a separable Hilbert space familiar from quantum mechanics, standardly called l^2. This includes the probability functions defined on S, in the same way as before, and the scalar product is well defined.

The definitions we gave above for the finite case carry over naturally. A hyperplane is a set of vectors orthogonal to a given vector, and equivalently, the set of vectors x such that for a certain vector y, (x,y) = 0. A hyperplane H divides the space into two halfspaces, $H^+ = \{x: (x,y) \geq 0\}$ and $H^- = \{x: (x,y) \leq 0\}$.

Although S is now allowed to be countably infinite, we continue to require that $\boldsymbol{R}$ be finite, so its convex hull is a polytope with the members of $\boldsymbol{R}$ its vertices. Since a subspace of l^2 is convex, all of this hull, together with the prior $\boldsymbol{p}$, is a subset of finite-dimensional subspace. Nothing in the argument in Theorem 4 required reference to the properties of the ambient space, so it applies entirely, without change, to $\boldsymbol{p}$ and $\boldsymbol{R}$ in this:

Funding: Halpern's work was supported in part by the National Science Foundation (NSF) under grants IIS-0911036 and CCF-1214844, the Air Force Office of Sponsored Research (AFOSR) under grant FA9550-08-1-0438, the Army Research Office (ARO) under grant W911NF-14-1-0017, and by the Department of Defence Multidisciplinary University Research Initiative (MURI) program administered by AFOSR under grant FA9550-12-1-0040.

Acknowledgements: We thank the referees for their detailed reading of the paper and many useful comments, which resulted in significant improvements. Van Fraassen thanks Simon Kochen for a helpful discussion of issues about convexity in infinite-dimensional spaces.

REFERENCES

Diaconis, P. and S. L. Zabell [1982]. 'Updating Subjective Probability'. *Journal of the American Statistical Association* 77: 822-830.

Dubins, L. E. [1975]. 'Finitely Additive Conditional Probabilities, Conglomerability and Disintegrations'. *The Annals of Probability* 3: 89-99.

Friedman, K.. [1973]. 'Replies to Tribus and Motroni and to Gage and Hestenes'. *Journal of Statistical Physics* 9: 265-269.

Grove, A. and J. Y. Halpern [1997]. 'Probability Update: Conditioning versus Cross-Entropy'. *Proceedings of the 13th Conference on Uncertainty in AI*: 208–214, Providence, Rhode Island.

Gruber, P. M. [2007]. *Convex and Discrete Geometry.* Berlin: Springer Verlag.

Grünwald, P. D. and J. Y. Halpern [2003] "Updating Probabilities". *Journal of Artificial Intelligence Research* 19:243-278.

Hartmann, S. and S. R. Rad [2012]. 'Updating on Conditionals = Kullback-Leibler + Causal Structure', presented at the *Philosophy of Science Association Biennial Conference*, San Diego, USA.

Jaynes, E. T. [1957]. 'Information Theory and Statistical Mechanics', *Physical Review* 106:4, pp. 620-630.

Jaynes, E. T. [2003]. *Probability Theory: The Logic of Science*, Cambridge University Press.

Jeffrey, R [1968]. 'Probable Knowledge', pp. l66-l80, 189-190 in I. Lakatos (ed.) *The Problem of Inductive Logic*. North Holland.

Luenberger, D. G. [1969]. *Optimization By Vector Space Methods.* New York: Wiley.

Seidenfeld, T. [1987]. 'Entropy and Uncertainty', 259-287 in I. B. MacNeill and G. J. Umphrey (eds.), *Foundations of Statistical Inference*. Dordrecht: Reidel Publishing Company.

Skyrms, B. [1980]. 'Higher Order Degrees of Belief,' pp. 109-138 in Hugh Mellor (ed.) *Prospects for Pragmatism.* Cambridge: Cambridge University Press.

Skyrms, B. [1985]. 'Maximum Entropy Inference as a Special Case Of Conditionalization'. *Synthese* 63: 55-74.

Skyrms, B. [1987]. 'Updating, Supposing, and MAXENT'. *Theory and Decision* 22: 225-246.

Skyrms, B. [2013]. *From Zeno to Arbitrage: Essays on Quantity, Coherence, and Induction*. New York: Oxford University Press.

Teller, P. [1976] "Conditionalization, Observation, and Change of Preference." In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science* (vol. I**),** eds. W.L. Harper and C.A. Hooker, 205-253**.** Dordrecht: Reidel.

Teller, P. and A. Fine. [1975] "A characterization of conditional probability". *Mathematics Magazine* 48: 267-270.

Tribus, M. and H. Motroni. [1972] 'Comments on the paper "Jaynes' Maximum Entropy Prescription and Probability Theory"'. *Journal of Statistical Physics* 4: 227 -228.

van Fraassen, B. C. [1981]. 'A Problem for Relative Information Minimizers in Probability Kinematics', *British Journal for the Philos. of Science* 32, 375-379.

van Fraassen, B. C., R. I. G. Hughes, and G. Harman [1986]. 'A Problem for Relative Information Minimizers in Probability Kinematics, Continued', *British Journal for the Philosophy of Science* 37: 453-475.

van Fraassen, B. C. [1995]. 'Belief and the Problem of Ulysses and the Sirens', *Philosophical Studies* **77**: 7–37.

van Fraassen, B. C. [1999]. 'Conditionalization, a New Argument For', *Topoi* 18: 93-96. Available: http://www.princeton.edu/~fraassen/abstract/mss.html

Williamson, J. [2011]. 'Objective Bayesianism, Bayesian Conditionalization, and Voluntarism', *Synthese* 178: 67-85.

ENDNOTES

[1] Department of Philosophy, San Francisco State University, San Francisco, CA 94132; fraassen@princeton.edu.

[2] Department of Computer Science, Cornell University, Ithaca, NY 14850; halpern@cs.cornell.edu

[3] For example, the strong law of large numbers implies that if G is a finitely generated field of subsets of S, and **p** a probability function defined on G, then there exists a countable sequence **s** of members of S such that for each element E of G, **p**(E) equals the limit of the relative frequency of E in **s**.

[4] If there is only one possible posterior we take it that this would correspond to something like a test of which the outcome is certain beforehand. In that case the prior should not change, and the posterior must be the same as the prior.

[5] For prior results concerning the conditions under which a single prior to posterior shift as involving a 'hidden' conditionalization see Diaconis and Zabell 1982; Skyrms 1980.

[6] The policy of conditionalization satisfies this criterion. But our focus here is on the more interesting question of how the criterion applies to updating policies allowing for a larger variety of inputs, and a less restrictive specification of allowable posteriors in response.

[7] We include here infinite (countable) combinations as convex, when the coefficients are real, positive, and sum to 1.

[8] This is deliberately formulated in such a way that it would apply also to interval-valued probability, but here we focus solely on the sharp probability case.

[9] For the special case in which the possible posteriors each assign 1 to distinct cells of a single partition, the result on conglomerability in Theorem 1 of Dubins 1975 is closely related to the following theorems. (Thanks to Teddy Seidenfeld for this reference.)

[10] We would have a much simpler proof, using the *Separating Hyperplane Theorem* (Gruber [2007]), p. 59; Luenberger [1969], pp. 133-4) if we did not required the probability p to lie strictly insider the interval spanned by $E_{p'}[g]$ for p' in **R**. Our proof is for the case where **S** and **R** are finite; we do not know if the result continues to hold if **R** is infinite (but see the Appendix).

[11] For the special case in which the posteriors have disjoint support, that was proved in [van Fraassen 1999]. The argument there does not apply to the general case.

[12] In fact, since [***R***] is part of the convex set of all probability functions on this space, it is bounded (all members have norm ≤ 1).