Tracking probabilistic truths: a logic for statistical learning

Baltag, A.; Rad, S.R.; Smets, S.

DOI
10.1007/s11229-021-03193-6

Publication date
2021

Document Version
Final published version

Published in
Synthese

Synthese (2021) 199:9041–9087
https://doi.org/10.1007/s11229-021-03193-6

ORIGINAL RESEARCH

Tracking probabilistic truths: a logic for statistical learning

Alexandru Baltag[1] · Soroush Rafiee Rad[1,2] · Sonja Smets[1,3]

Received: 23 November 2020 / Accepted: 29 April 2021 / Published online: 8 September 2021

Abstract
We propose a new model for forming and revising beliefs about unknown probabilities. To go beyond what is known with certainty and represent the agent's *beliefs* about probability, we consider a plausibility map, associating to each possible distribution a plausibility ranking. Beliefs are defined as in Belief Revision Theory, in terms of truth in the most plausible worlds (or more generally, truth in all the worlds that are plausible enough). We consider two forms of conditioning or belief update, corresponding to the acquisition of two types of information: (1) learning observable evidence obtained by repeated sampling from the unknown distribution; and (2) learning higher-order information about the distribution. The first changes only the plausibility map (via a 'plausibilistic' version of Bayes' Rule), but leaves the given set of possible distributions essentially unchanged; the second rules out some distributions, thus shrinking the set of possibilities, without changing their plausibility ordering.. We look at stability of beliefs under either of these types of learning, defining two related notions (safe belief and statistical knowledge), as well as a measure of the verisimilitude of a given plausibility model. We prove a number of convergence results, showing how our agent's beliefs track the true probability after repeated sampling, and how she eventually gains in a sense (statistical) knowledge of that true probability. Finally, we sketch the contours of a dynamic doxastic logic for statistical learning.

Keywords Radical uncertainty · Imprecise probabilities · Plausibility models · Statistical learning · Multinomial distribution · Belief revision theory · Doxastic logic · Formal epistemology

Mathematics Subject Classification 03B42 · 03B48 · 03B60

This article belongs to the topical collection on Approaching Probabilistic Truths, edited by Theo Kuipers, Ilkka Niiniluoto, and Gustavo Cevolani.

Extended author information available on the last page of the article

1 Introduction

The goal of this paper is to propose a new model for *learning a probabilistic distribution*, in situations that are commonly characterized as those of "radical uncertainty" (Walley 1996) or "Knightian uncertainty" (Cerreia-Vioglio et al. 2013). The most widespread model for these situations uses *imprecise probabilities*, i.e. *sets* of probability distributions. As an example, consider an urn full of marbles, coloured red, green, and blue, but with an unknown distribution. What is then the probability of drawing a red marble? In such cases, when the agent's information is not enough to determine the true distribution, she is typically left with a large (possibly infinite) set of possible probability assignments. If she never goes beyond what she knows, then her only 'rational' answer should be "I don't know": she is in a state of *ambiguity*, and she should simply consider possible *all* distributions that are consistent with her background knowledge and observed evidence. This type of over-cautious rationality, resembling the famous paradox of "Buridan's ass", is not of much help in dealing with practical decision problems.

Our model allows the agent to go beyond what she knows with certainty, by forming *rational qualitative beliefs* about the unknown distribution, beliefs based on the inherent plausibility of each possible distribution. For this, we assume the agent is endowed with an initial *plausibility map*, assigning real numbers to the possible distributions. The plausibility map encodes the agent's background beliefs and a priori assumptions about the world. For instance, an agent who assumes the Principle of Indifference (Williamson 2013; Hájek 2019) will use Shannon entropy as her plausibility function, thus initially believing that the distribution is *the most non-informative* one (in the given set of possibilities). On the other hand, an agent assuming a Normality or 'Averageness' Principle, will use closeness to the Center of Mass or the barycenter (Paris 1994) as her plausibility measure, thus starting with a belief in the *most typical* distribution, i.e. the one that is the most representative for the given set of distributions. Finally, an agent who assumes some form of Ockham's Razor will use as plausibility some measure of simplicity (Kelly 2008), thus her prior belief will focus on the *simplest* distribution(s).

Our agent forms beliefs by using the standard definition of qualitative belief in Logic and Belief Revision Theory, in terms of *plausibility maximization* (Board 2004; Baltag and Smets 2008b): she believes the most plausible distribution(s). More precisely, we equate "belief" with "truth in all the worlds that are plausible enough": P is believed iff there exists some distribution μ s.t. P is true in all distributions that are at least as plausible as μ. In particular, "belief" coincides with *truth in all the most plausible worlds*, whenever such most plausible worlds/distributions exist. As a consequence, all the usual KD45 axioms of doxastic logic will be valid in our framework.

Note that, although our plausibility map assigns real values to probability distributions, this account is essentially different from the ones using so-called "second-order probabilities" (i.e. probability distributions defined on the given set of probability distributions) (Gaifman and Snir 1982; Gaifman 2016). Plausibility values are only relevant in so far as they induce a qualitative order on distributions. In contrast to probability, plausibility is *not cumulative* (in the sense that the low-plausibility alternatives do not add up to form more plausible sets of alternatives), and as a result only

higher-ranking distributions 'beat' lower-ranking ones; in case that some distributions have the *highest* plausibility, they are the only ones of any relevance for beliefs.

Our model is not just a way to "rationally" select a Bayesian prior, but it also comes with a rational method for *revising beliefs* in the face of new evidence. In fact, it can deal with *two types of new information*: first-order evidence gathered by repeated *sampling* from the (unknown) distribution; and higher-order information about the distribution itself, coming in the form of a *set* of possible distributions (often defined by a set of linear inequality constraints on that distribution). To see the difference between the two types of new evidence, take for instance the example of a coin. As it is well-known, any finite sequence of Heads and Tails is consistent with all possible non-extremal biases of the coin. As such, any number of finite repeated samples *will not* shrink the set of possible biases, though they may increase the plausibility of some biases. Thus this type of information changes only the plausibility map but leaves the given set of distributions essentially unchanged (except for the elimination of some extremal distributions, that assigned probability 0 to the observed sample). The second type of information, on the other hand, shrinks the set of measures, while keeping their relative plausibility ranking. For instance, learning that the coin has a bias towards Tail (e.g. by weighing the coin, or receiving a communication in this sense from the coin's manufacturer) eliminates all distributions that assign a higher probability to Heads. It is important to notice, however, that even with higher-order information, it is hardly ever the case that the distribution under consideration is fully specified. In our coin example, a known bias towards Tails will still leave an infinite set of possible biases consistent. Even a good measurement by weighting will leave open a whole interval of possible biases. In this sense, a combination of observations and higher-order information will *not* in general allow the agent to come to *know* the correct distribution, in the standard ('infallible') sense in which the term knowledge is used in doxastic and epistemic logics. Instead, it may eventually allow her to come to *believe* the true probability (at least, with a high degree of accuracy). This belief may even stabilize, to such a degree that it approaches the 'softer', defeasible notion of 'knowledge', which is the main focus in Epistemology (Lehrer 1990; Stalnaker 1996; Rott 2004) and (inductive) Learning Theory (Gold 1967; Baltag et al. 2019a). This *convergence in belief* and the resulting *acquisition of statistical knowledge* is what we aim to capture in this paper.

Our mechanism for belief revision with sampling evidence is non-Bayesian (and also different from AGM belief revision), though it incorporates a "plausibilistic" version of Bayes' Rule. Instead of updating her prior belief according to this rule (and disregarding all other possible distributions), the agent keeps all possibilities in store and *revises instead, her plausibility ranking*, using a non-probabilistic analogue of Bayes' Rule. After that, her new belief will be formed in a similar way to her initial belief: by maximizing her (new) plausibility. The outcome is different from simply performing a Bayesian update on the 'prior': qualitative jumps are possible, leading to abandoning "wrong" conjectures in a non-monotonic way. This results in a *faster* convergence-in-belief to the true probability in *less restrictive conditions* than the

usual Savage-style convergence through repeated Bayesian updating (Edwards et al. 1963; Savage 1954).[1]

The second type of evidence (higher-order information about the distribution) induces a more familiar kind of update: the distributions that do not satisfy the new information (typically given in the form of linear inequalities) are simply eliminated, then beliefs are formed as before by focusing on the most plausible remaining distributions. This form of revision is known as AGM *conditioning* in Belief Revision Theory (Alchourrón et al. 1985), and as *update* or "public announcement" in Logic (Baltag and Renne 2016; van Ditmarsch et al. 2007), and satisfies all the standard AGM axioms.[2]

The fact that in our setting there are two types of updates should not be so surprising. It is related to the fact that our static framework consists of two different semantic ingredients, capturing two types of information: the *plausibility* map (encoding the agent's *beliefs and conditional beliefs*, defeasible forms of knowledge, etc) and the *set* of possible distributions (encoding the agent's *infallible knowledge*, her 'hard information' about the correct distribution). Correspondingly, the first type of update directly affects the agent's beliefs (by changing the plausibility in the view of the sampling results), and only indirectly her knowledge (since e.g. she knows her new beliefs). Dually, the second type of update directly affects the agent's knowledge (by reducing the set of possibilities), and only indirectly her beliefs (by restricting the plausibility map to the new set).

By allowing two forms of learning, one having a Bayesian-statistical flavor and the other having a logical-AGM flavor (Alchourrón et al. 1985; Darwiche and Pearl 1997), our framework combines logical and statistical reasoning in a unified setting. In this sense, it fits within the recent trend towards a unification of logic and probability, see e.g. Leitgeb (2017). In particular, the fact that conditioning on sampling evidence is non-AGM is in fact *essential* for the successful learning of the true probability from repeated sampling: since every sample is logically consistent with every non-extremal distribution, an AGM learner (obeying the principle of Rational Monotonicity[3]) would typically never change her initial beliefs about the true distribution after any number of

[1] In contrast to Savage's theorem, our update ensures convergence even in the case that the initial set of possible distributions is infinite (indeed, even in the case we start with the uncountable set of *all* distributions). Moreover, in the finite case (where Savage's result does apply), our update is guaranteed to converge in finitely many steps, while Savage's theorem only ensures convergence in the limit.

[2] We should note that there have been several proposals in the literature for AGM-compatible processes of iterated belief revision to remedy the inadequacy of AGM postulates to correctly capture the process of belief change from repeated observations, see for example Booth and Meyer (2006), Darwiche and Pearl (1997), Konieczny and Perez (2000) and Nayak (1994). In fact, the propositional conditionalization in our paper, like its older qualitative versions in Game Theory (Board 2004) and Dynamic Epistemic Logic (Baltag and Smets 2008c; Baltag and Renne 2016; van Ditmarsch et al. 2007), is an instance of the iterated revision operation of Darwiche and Pearl (1997): indeed, both the prior information state before revision and the revised information state are *plausibility models* (i.e. what (Darwiche and Pearl 1997) calls "epistemic states"), rather than theories or belief bases (i.e. propositions or sets of propositions). Still, we follow the usage in the epistemological (Kelly 2014; Kelly et al. 1995, 1998) and dynamic-epistemic logic literature (Baltag et al. 2016; Baltag and Renne 2016; van Ditmarsch et al. 2007) in calling this operation AGM conditioning.

[3] Rational Monotonicity says that, if the prior beliefs are consistent with the incoming evidence, then they continue to be believed after revising with the evidence. Rational Monotonicity is not actually one of the AGM postulates, but it is specific to logical presentations of this framework in terms of conditional

samples! The same applies to all the generalizations of AGM conditioning that retain Rational Monotonicity, e.g. the ones proposed by Darwiche and Pearl (1997), or by Konieczny and Perez (2000).

A preliminary version of this paper was presented at TARK 2019, and an abstract appeared in the online proceedings (Baltag et al. 2019). Our current article is the extended, journal version of that work, though with many major changes: improvements of the basic setting, the formalization and study of new epistemic notions (e.g. safe belief of a distribution, statistical knowledge, distance-from-the-truth), and a number of new convergence results. The plan of the paper is as follows. We start by reviewing in Sect. 2 some basic notions, results, and examples on probability distributions. In Sect. 3, we define our main setting (probabilistic plausibility models), consider a number of standard examples, define in this setting the notions of belief and (infallible) knowledge, and study their logical properties. In Sect. 4, we move to conditional beliefs, defining our two forms of conditionalization, and use them to explore belief dynamics (as captured by our two types of model updates). In Sect. 5, we look at notions of doxastic stability, defining a weaker form of stability ("safe belief"), followed by a stronger form ("statistical knowledge"), and investigating their properties and their connection to a notion of verisimilitude (or "distance from the truth"). In Sect. 6, we present and prove our main results on doxastic convergence to the true probability. Finally, in Sect. 7 we briefly sketch the contours of a dynamic doxastic logic for statistical learning, and in Sect. 8 we end with some concluding remarks and a brief comparison with other approaches to the same problem.

2 Preliminaries and notation

Throughout this paper, we fix a finite set $O = \{o_1, \ldots, o_n\}$ of possible *observations*, or '(elementary) outcomes'.[4] Let

$$M_O := \left\{ \mu \in [0,1]^O \mid \sum_{o \in O} \mu(o) = 1 \right\}$$

be the set of probability mass functions on O, which we identify with the corresponding probability functions on $\mathcal{P}(O)$. The sets of distributions $P \in \mathcal{P}(M_O)$ will be called *propositions*. Let

$$\Omega = O^\infty := \{\omega \mid \omega : \mathbb{N}\backslash\{0\} \to O\}$$

be the set of infinite sequences from O, which we shall refer to as *observation streams*. Each such stream $\omega = (\omega_1, \ldots, \omega_n, \ldots)$ represents a possible history of future sampling from an unknown distribution. For any $\omega \in \Omega$ and $i \in \mathbb{N}\backslash\{0\}$, we write ω_i for the i-th component of ω, and $\omega^{\leq i}$ for its initial segment of length i, i.e. the sequence

beliefs (Board 2004; Baltag and Smets 2008b), and it is equivalent to a combination of AGM axioms of Inclusion/Subexpansion and Vacuity/Superexpansion.

[4] Intuitively, these are the possible outcomes of sampling or of some other possible type of experimentation.

$\omega^{\leq i} := (\omega_1, \ldots, \omega_i)$ consisting of the first i components of ω. Similarly, we put $\omega^{>i} := (\omega_{i+1}, \ldots, \omega_n, \ldots)$ for the infinite "tail" of ω that follows the i-th observation. In particular, $\omega^{\leq 0} := \lambda = ()$ is the empty sequence, and $\omega^{>0} = \omega$. We denote by

$$O^* = \{(\omega_1, \ldots, \omega_i) | \, i \geq 0, \omega_1, \ldots, \omega_i \in O\}$$

the set of all finite sequences of observations. For each $o \in O$ we define the sets o^j to be the basic cylinders

$$o^j = \{\omega \in \Omega \, | \, \omega_j = o\} \subseteq \Omega.$$

These cylinders correspond to individual observations of evidence sampled from the unknown distributions. Let $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ be the σ-algebra of subsets of Ω generated by the cylinders (algebra obtained by closing the family of basic cylinders under complementation and countable unions). Every probability distribution $\mu \in M_O$ induces a unique multinomial probability distribution over $(\Omega, \mathcal{A})$, also denoted by μ, and obtained by first setting

$$\mu(o^j) = \mu(o)$$

then extending this to all of $\mathcal{A}$ using independence, additivity and continuity. Let $\mathcal{E} \subseteq \mathcal{A}$ be the family of sets obtained by closing the family of basic cylinders only under complementation and *finite* unions. The sets $e \in \mathcal{E}$ are called *observable events* (or just 'events', for short).[5] It is easy to see that every event $e \in \mathcal{E}$ can be written as a *finite disjoint union of finite intersections of basic cylinders*. In particular, for each finite sequence of observations $\omega^{\leq i} = (\omega_1, \ldots, \omega_i) \in O^*$, we denote by $[\omega^{\leq i}] = [\omega_1, \ldots, \omega_i]$ the corresponding event of observing this sequence by sampling, i.e. the event given by

$$[\omega^{\leq i}] = [\omega_1, \ldots, \omega_i] := \{\omega' \in \Omega : | \, \omega'_j = \omega_j \text{ for all } j \leq i\} = \bigcap_{j=1}^{i} \omega^j_j$$

Example 1 Let $O = \{H, T\}$ be the possible outcomes of a coin toss. Then Ω will be streams of *Heads* and *Tails* representing infinite tosses of the coin, e.g. HTTHHH.... And H^j (res. T^j) will be the set of streams of observations in which the j-th toss of the coin will land Heads up (res. Tails up). The set M_O will be the set of possible biases of the coin.

Example 2 Let $O = \{R, B, G\}$ be the possible outcomes for a draw from an urn filled with marbles, coloured Red (R), Blue (B) and Green (G). Then M_O will be the set of all possible distributions of coloured marbles in the urn, Ω will be the set of infinite streams of R, B and G (representing infinite draws from the urn), and R^j (res. B^j or

[5] In the literature, the term 'event' is also used for all the members of the σ-algebra $\mathcal{A}$, but this family includes unobservable events, such as a coin falling Heads up infinitely many times.

G^j) will be the set of streams of draws in which the j-th draw is a Red (res. Blue or Green) marble.

Standard topology on M_O Notice that a probability function $\mu \in M_O$, defined over the set $O = \{o_1, \ldots, o_n\}$, can be identified with an n-dimensional vector $(\mu(o_1), \ldots, \mu(o_n))$, corresponding to the probabilities assigned to each o_i respectively. Let $\mathcal{D}_O := \{\mathbf{x} \in [0,1]^n \mid \sum x_i = 1\}$, then every $\mu \in M_O$ can be identified with the point $\bar{} \in \mathcal{D}_O \subset [0,1]^n$. Thus probability functions in M_O live in the vector space $[0,1]^n$. In the other direction every $\mathbf{x} \in \mathcal{D}_O$ defines a probability function x on O by setting $x(o_i) = \mathbf{x}_i$. This gives a one to one correspondence between M_O and $\mathcal{D}_O$. There are various metric distances that can be defined on the space of probability measures over a (finite) set O, many of which are known to induce the same topology. Here we will consider the *standard topology* of $[0,1]^n$, induced by the Euclidean metric: for $\mathbf{x}, \mathbf{y} \in [0,1]^n$, put $d(\mathbf{x}, \mathbf{y}) := \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$; a basis for the standard topology is given by the family of all *open balls* $\mathcal{B}_\varepsilon(\mathbf{x})$ *centred at some point* $\mathbf{x} \in \mathbb{R}^n$ *with radius* $\varepsilon > 0$; where

$$\mathcal{B}_\varepsilon(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^n \mid d(\mathbf{x}, \mathbf{y}) < \varepsilon\}.$$

We will make use of the following well-known facts:

Proposition 1 *For any finite set* O*, the set* M_O *of probability mass functions on* O *is compact in the standard topology.*

Proof Notice that the set $\{\mathbf{x} \in [0,1]^n \mid \sum_{i=1}^n x_i = 1\}$ is compact in $\mathbb{R}^n$. □

Proposition 2 *Let* X, Y *be compact topological spaces,* $Z \subseteq X$ *and* $f : X \to Y$

(1) *Every closed subset of* X *is compact.*
(2) *If* f *is continuous, then* $f(X)$ *is compact.*
(3) *If* Z *is compact then it is closed and bounded.*

Proof See Hunter (2012), Theorem 1.40 and Proposition 1.41. □

Lemma 1 *For each event* $e \in \mathcal{E}$*, the function* $F_e : M_O \to [0,1]$*, defined as* $F_e(\mu) := \mu(e)$*, is continuous.*

Proof This can be verified, using the above-mentioned fact that every event is a finite disjoint union of finite intersections of basic cylinders. The proof is by induction on the structure of this representation. The conclusion is immediate when $e = o^j$ is a basic cylinder: given any $\varepsilon > 0$, we can take $\delta := \varepsilon$, and then, for all $\mu, \nu \in M_O$ with $d(\mu, \nu) < \delta$, we have $|F_e(\mu) - F_e(\nu)| = |\mu(e) - \nu(e)| = |\mu(o) - \nu(o)| \leq \sqrt{\sum_{i=1}^n (\mu(o_i) - \nu(o_i))^2} = d(\mu, \nu) < \delta = \varepsilon$. This can be extended to finite intersections of basic cylinders, by noting that if $e = \bigcap_{k=1}^m \omega_k^{j_k}$ is such a finite intersection with all j_k distinct,[6] then by independence we have $F_e = \prod_{k=1}^m F_{\omega_k^{j_k}}$, and then using the

[6] If $j_k = j_q$ and $\omega_k \neq \omega_q$ for some $j \neq q$, then the intersection is empty; while if $j_k = j_q$ and $\omega_k = \omega_q$ (for $j \neq q$), then we have $\omega_k^{j_k} = \omega_q^{j_q}$, so one of the two terms is redundant and can be eliminated from the representation.

fact that a finite product of continuous functions is continuous. Finally, we can extend to disjoint unions of finite intersections of basic cylinders, noting that if $e = \bigcup_{i=1}^{m} e_i$ is a disjoint union of events (with $e_i \cap e_j = \emptyset$ for $i \neq j$), then by additivity we have $F_e = \sum_{i=1}^{m} F_{e_i}$, and then using the fact that a finite sum of continuous functions is continuous. □

Proposition 3 *Every continuous function $f : X \to \mathbb{R}$ on a compact topological space X is bounded, and it attains its supremum (i.e., it has a maximum value).*

Proof See Hunter (2012), Theorem 7.35. □

Theorem 1 (Hein-Cantor) *Let M, N be two metric spaces and $f : M \to N$ be continuous. If M is compact then f is uniformly continuous.*

Proof See Rudin (1953). □

Before presenting our framework, we need one more technical lemma that will prove useful in the proof of our convergence Theorem 1.

Lemma 1 *For $0 < p_1, \ldots, p_n \leq 1$ with $\sum p_i = 1$, the function $f(\mathbf{x}) = \Pi_{i=1}^{n} x_i^{p_i}$ has $\mathbf{x} = \mathbf{p}$ as its unique maximizer on M_O.*

Proof First we notice that $f(\mathbf{x}) \geq 0$ on $M_O = \{\mathbf{z} \in [0,1]^n \mid \sum z_i = 1\}$ and by Propositions 1 and 3 f has a maximum value on M_O. But note that $f(\mathbf{z}) = 0$ for any point $\mathbf{z} \in M_O$ having some zero coordinate $z_i = 0$ (for any $i \leq n$). Hence, f reaches its maximum on $D = (0,1]^n \cup M_O = \{\mathbf{z} \in (0,1]^n \mid \sum z_i = 1\}$.

To prove the lemma, we will show that $\log(f(\mathbf{x}))$ has $\mathbf{x} = \mathbf{p}$ as its unique maximizer on D. The conclusion will then follow from noticing that $f(x) \geq 0$ and the monotonicity of log function on $\mathbb{R}^+$. To maximise $\log(f(\mathbf{x}))$ subject to condition $\sum_i x_i = 1$, we use Lagrange multiplier methods: let

$$G(\mathbf{x}) = \log(f(\mathbf{x})) - \lambda \left(\sum_{i=1}^{n} x_i - 1 \right) = \sum_{i=1}^{n} p_i \log(x_i) - \lambda \left(\sum_{i=1}^{n} x_i - 1 \right).$$

Setting partial derivatives of G equal to zero we get,

$$\frac{\partial G(\mathbf{x})}{\partial x_i} = \frac{p_i}{x_i} - \lambda = 0$$

which gives $p_i = \lambda x_i$. Inserting this in the condition $\sum_i p_i = 1$ we get $\lambda \sum_i x_i = 1$ and using $\sum_i x_i = 1$ we get $\lambda = 1$ and thus $x_i = p_i$. Since f has a maximum on this domain and the Lagrange multiplier method gives a necessary condition for the maximum, any point $\mathbf{x}$ that maximises f should satisfy the condition $x_i = p_i$ and thus $\mathbf{p}$ is the unique maximiser for f. □

3 Probabilistic plausibility models

In this section, we introduce and exemplify our basic framework for dealing with radical uncertainty.

Definition 1 (Plausibility measures) A *plausibility 'measure' (on K)* is a continuous function $pla : K \to [0, \infty)$, whose domain is some closed set of distributions $K = \overline{K} \subseteq M_O$. Given a plausibility measure on K, we can extend it to a map[7] on *propositions* (sets of distributions) $P \subseteq M_O$, by putting

$$pla(P) \;:=\; sup\{pla(\mu) \mid \mu \in P \cap K\}$$

Similarly, we can extend it to *distribution-event pairs* $(\mu, e) \in K \times \mathcal{E}$, by putting:

$$pla(\mu, e) \;:=\; pla(\mu) \cdot \mu(e),$$

and further extend this to *proposition-event pairs* $(P, e) \in \mathcal{P}(M_O) \times \mathcal{E}$, by putting

$$pla(P, e) \;:=\; sup\{pla(\mu, e)) \mid \mu \in P\} = sup\{pla(\mu) \cdot \mu(e) \mid \mu \in P \cap K\}$$

These last two maps give us a way of assessing the *joint plausibility* of having true distribution μ (or true proposition P) *and* observing event e.

It seems apt at this point to emphasize again that events $e \in \mathcal{E} \subseteq \mathcal{A}$ in our setting are intended to capture observable events in multinomial experiments. The successive observations ω_i in a finite sampling sequence $\omega^{\leq i} = (\omega_1, \ldots, \omega_i)$ are thus regarded as outcomes of i independent and identically distributed trials as in Examples 1 and 2 . In the same manner, $\mu(e)$ encodes the probability assigned to e by the unique multinomial probability distribution induced by each $\mu \in M_O$ on $(\Omega, \mathcal{A})$ (which by a slight abuse of notation we also denote by μ).

Definition 2 [Plausibility models] A *(probabilistic) plausibility model* is a structure $\mathbf{M} = (M, pla)$ where $M \neq \emptyset$ is a non-empty subset of M_O, called the *set of 'possible distributions'*, and $pla : \overline{M} \to [0, 1]$ is a plausibility measure on the closure $\overline{M}$, called *probabilistic plausibility ranking map* (or just plausibility map, for short), and required to satisfy two additional conditions: (1) possible distributions have positive plausibility rank, i.e. $pla(\mu) > 0$ for all $\mu \in M$; (2) $pla(M) = 1$, or equivalently the maximum plausibility value on $\overline{M}$ is 1.[8]

[7] Using systematic ambiguity, we also denote this map by pla.

[8] The equivalence between these conditions is easily seen if we note that, by the continuity of plausibility measures, we have $pla(M) = sup\{pla(\mu) \mid \mu \in M\} = max\{pla(\mu) \mid \mu \in \overline{M}\} = pla(\overline{M})$. Note that $pla(M) = 1$ implies that *there exist possible distributions (in M) with plausibility arbitrarily close (or equal) to* 1.

The plausibility map induces a total preorder[9] $\leq^{\mathbf{M}}$ on the possible distributions in M, called the *plausibility ranking order*, and given by putting for all $\mu, \nu \in M$:

$$\mu \leq^{\mathbf{M}} \nu \text{ iff } pla(\mu) \leq pla(\nu).$$

For every real number $\delta \in [0, 1]$, we put $M^{\delta} := \{\mu \in M \mid pla(\mu) \geq \delta\}$ for the set of all distributions in M that have plausibility rank at least δ. A *(probabilistic) Grove sphere* is a non-empty set of the form $M^{\delta} \neq \emptyset$.[10] It is easy to see that the family of all Grove spheres $\mathcal{S} := \{M^{\delta} \mid \delta \in [0, 1], M^{\delta} \neq \emptyset\}$ is *nested* (i.e. totally ordered by inclusion: in fact, for $\delta \geq \varepsilon$ we have $M^{\delta} \supseteq M^{\varepsilon}$), and *exhaustive* (i.e. $M = \bigcup \mathcal{S}$).

The plausibility map pla attains its supremum (1) on M if and only if there exists a *smallest Grove sphere*, given by the set

$$Max(M) := M^{1} = \{\nu \in M \mid pla(\nu) = 1\} = \{\nu \in M \mid \nu \geq^{\mathbf{M}} \nu' \text{ for all } \nu' \in M\}$$

of all *maximizers* of the function pla on M.

The *plausibility* $pla(e)$ *of an event* e *in the model* $\mathbf{M} = (M, pla)$ is defined as the joint plausibility $pla(M, e)$:

$$pla(e) \;:=\; pla(M, e) = sup\{pla(\mu) \cdot \mu(e) \mid \mu \in M\}$$

A plausibility model $\mathbf{M} = (M, pla)$ is said to be *closed* if the set M of possible distributions is closed (in the standard topology on M_O). The model is said to be *convex* if the set M is convex (i.e. $\alpha \cdot \mu + (1 - \alpha) \cdot \nu \in M$ for all $\mu, \nu \in M$ and all $\alpha \in [0, 1]$).

The difference between plausibility measures and (the special case of) plausibility ranking maps is a plausibilistic analogue of the difference between measures in Measure Theory and (the special case of) probability functions. Although conditions (1) and (2) in the definition of plausibility maps may appear very restrictive at first sight, they do not in fact restrict the generality of our plausibility ranking order: the next example shows that *any plausibility measure* can be used to define plausibility ranking maps.

Generic example: plausibility-generating measures Let $pla : K \to [0, \infty]$ be any plausibility measure with $dom(pla) = K = \overline{K} \subseteq M_O$. Then pla induces a plausibility model $\mathbf{M} = (M, pla^{M})$ on each non-empty subset $M \subseteq \{\mu \in K \mid pla(\mu) \neq 0\}$, with the plausibility map pla^{M} given by *renormalizing pla to M*, i.e. putting

$$pla^{M}(\mu) \;:=\; \frac{pla(\mu)}{pla(M)} = \frac{pla(\mu)}{sup\{pla(\nu) \mid \nu \in M\}} = \frac{pla(\mu)}{max\{pla(\nu) \mid \nu \in \overline{M}\}}$$

[9] A total preorder on M is a relation $\leq \subseteq M \times M$, which is reflexive ($\mu \leq \mu$ holds for every $\mu \in M$), transitive ($\mu \leq \nu \leq \rho$ implies $\mu \leq \rho$) and connected (either $\mu \leq \nu$ or $\nu \leq \mu$, or both) hold, for every pair $\mu, \nu \in M$.

[10] Note that we have $M^{\delta} \neq \emptyset$ for every $\delta < 1$, so these are always Grove spheres; $\delta = 1$ is the only value of δ for which M^{δ} may fail to be a sphere.

for all $\mu \in \overline{M}$. In this case, we say that the plausibility ranking map pla^M is *generated by the plausibility measure* pla. Note that the plausibility ranking order $\leq^{\mathbf{M}}$ induced by pla^M on M coincides with the order induced by the generating measure pla, i.e. we have:

$$\mu \leq^{\mathbf{M}} \mu' \text{ iff } pla(\mu) \leq pla(\mu').$$

A plausibility-generating measure pla is said to be *fully positive* whenever its domain $dom(pla) = M_O$ is the full set of all distributions, and its codomain is $(0, \infty)$ (i.e. $pla(\mu) > 0$ for all $\mu \in M_O$). This is a special case of great importance: *fully positive measures generate plausibility models on every non-empty set of distributions* $M \subseteq M_O$.

Interpretation In a plausibility model, the current set of possibilities M encodes an agent's *current epistemic state*, her "hard information" or *higher-level knowledge* about a given probabilistic distribution μ: all she knows for sure is that $\mu \in M$. The agent may have come to this prior knowledge due to some previously received information (either in the form of observations obtained by sampling or in the form of higher-level information about the mechanism underlying the unknown distribution). On the other hand, pla represents the agent's "soft information", her *current beliefs* (and conditional beliefs etc) about the unknown distribution, typically acquired by sampling. Unlike in probabilistic inference processes (Paris 1994)(but like in most concrete examples of such processes), this doesn't give only one (unconditional) belief, but a whole ranking of the distributions, in the form of a continuous function (which will give rise to a series of conditional beliefs): she considers the higher-ranked distributions to be *more plausible* than the lower-ranked ones. But, in contrast to knowledge, such soft information is not enough to exclude the less plausible distributions: the agent 'believes' that they are not the real distribution; but she doesn't know it for certain. The agent believes every proposition satisfied by all the "top" (most plausible) distributions: the ones having plausibility rank 1; or, if such top distributions don't exist, the agent will believe every proposition satisfied by all distributions that are "plausible enough": i.e. all above any given plausibility rank $1 - \varepsilon$ (for *any* $\varepsilon > 0$).

The above-defined extensions of the plausibility map have epistemic/doxastic significance: $pla(\mu, e)$ can be thought of as a way of assessing of *joint plausibility* of having true distribution μ *and* observing event e. Note the analogy with the formula for the joint probability of two events.[11] Similarly, $pla(P)$ gives us a way to assess the plausibility of a 'proposition': essentially, a set of distributions $P \subseteq M$ is only as plausible as *the most plausible* element of P (if such an element exists); or more generally P is at least as plausible as all its elements, but no more than (i.e. $pla(P)$ is the supremum of all plausibility ranks in P). Note now the analogy with, but also the difference from, probability: the role usually played by addition is played here by the supremum. With this notation, condition (2) on plausibility models (M, pla) can be restated simply as $pla(M) = 1$. Finally, $pla(P, e)$ combines the formulas for $pla(\mu, e)$ and for $pla(P)$ in the natural way, giving the *joint plausibility of having the*

[11] This analogy can be made more precise if we identify the *plausibility of an event e given a distribution* μ with the probability assigned by μ to e, i.e. put $pla(e|\mu) := \mu(e)$. Then the joint plausibility formula reads $pla(\mu, e) = pla(\mu) \cdot pla(e|\mu)$.

true distribution in P and observing event e. In particular, $pla(e) := pla(M, e)$ is a natural definition for the plausibility of the event e.

Differences between plausibility and probability Note the key differences between plausibility models and probabilistic models. First, unlike in the probabilistic case, maximal plausibility $pla(\mu) = 1$ does *not* mean certainty or full belief, but only *consistency with all the agent's beliefs*: the distributions μ with $pla(\mu) = 1$ are "doxastically possible", i.e. they satisfy every proposition believed by the agent. Second, the plausibility map does *not* obey Kolmogorov's additivity axiom: the plausibility *pla(P)* of a set is *not* the sum of plausibility ranks of its elements, but rather their supremum. This, together with the above normalization requirement (2), suggests that *the plausibilistic analogue of addition of probabilities is the operation of maximization* (or more generally, taking the supremum).

Models for experimental-based information *Closed models* characterize the situations in which *all prior knowledge about the distribution is based only on experimental evidence* about the mechanism underlying this distribution: e.g. measurements of the side weights or asymmetries of a coin or dice; opening each of a number of urns (from which an unknown one will be chosen for later sampling) and counting (or approximately estimating) the marbles of a given color in the urn, etc. In such contexts, it is indeed natural to assume that M is closed: if a distribution is a limit of possible distributions in M, then it is indistinguishable from M by any such experimental means, and hence it cannot be excluded from M.

In the case that the experimental evidence is based only on *measurements*, it is natural to assume more, namely that M is *both closed and convex*: measurements typically produce *interval estimates* $[a, b]$ for the probability $\mu(o)$ of each outcome. Indeed, such interval models are the ones most used when dealing with imprecise probabilities. More generally, the information obtained in this way may come in the form of *linear constraints* of the form $\sum_{i=1}^{n} a_i \mu(o_i) \geq c$ (with $a_1, \ldots, a_n, c \in Q$). Any finite set of such constraints gives a closed and convex set M of possible distributions.

One might wonder why do we permit distributions $\overline{M} \setminus M$ to have positive plausibility ranks, or even why do we take the whole closure $\overline{M}$ (instead of M) as the domain of the plausibility map. Given that the agent *knows for sure* that the true distribution lies within M, the distributions in $\overline{M} \setminus M$ are incompatible with the agent's hard information, so they are known to be 'impossible' in the view of this information. It would seem natural to require that $pla \equiv 0$ on $\overline{M} \setminus M$, or else just restrict the domain of *pla* to M. This can indeed be done *if M is closed*. But in general, the technical condition of continuity poses constraints on the plausibility ranks of distributions in the closure $\overline{M}$, which may force some $\mu \in \overline{M} \setminus M$ to have non-zero plausibility ranks. Even from a purely conceptual perspective, distributions in $\overline{M} \setminus M$ are in a sense "almost possible", since they are not distinguishable from the ones in M by any experimental means. Their epistemic impossibility is only due to higher-order, non-experimental information, and so it makes sense to take them into account. Moreover, it may be that such ideal limit-distributions may have a high inherent plausibility (despite being ruled out by the current information). In some cases, they may be inherently *more plausible* than the possible distributions. In such cases, these distributions would be in principle

believed on purely a priori grounds, though they are disbelieved (in fact known to be impossible) when the higher-level information is taken into account.[12]

The above intuitions about knowledge and belief can be made formal as follows:

Definition 3 [*Knowledge and belief*] We say that a proposition $P \subseteq M_O$ is *known* in the model $\mathbf{M} = (M, pla)$, and write $\mathbf{M} \models K(P)$, if *all possible distributions* are in P; i.e. if $M \subseteq P$.

We say that $P \subseteq M_O$ is *believed* in the model $\mathbf{M} = (M, pla)$, and write $\mathbf{M} \models B(P)$, if *all "plausible enough" distributions in M* are in P; i.e. iff there exists some $\mu \in M$ such that $\{\nu \in M \mid \nu \geq^{\mathbf{M}} \mu\} \subseteq P$. An equivalent definition can be given in terms of Grove spheres: *B(P)* holds in $\mathbf{M}$ iff P includes some Grove sphere; i.e. iff there exists $\delta \leq 1$ such that $\emptyset \neq M^{\delta} \subseteq P$; or, yet another equivalence: there exists $\varepsilon \geq 0$ such that $\emptyset \neq M^{1-\varepsilon} \subseteq P$.

Connections to belief revision theory Grove sphere models (in non-probabilistic form, consisting of possible worlds instead of distributions) form the standard semantic framework in Belief Revision Theory (Grove 1988). Plausibility models (again, in their non-probabilistic version) are well-known equivalent relational descriptions of sphere models, that are preferred in Dynamic Epistemic Logic (Baltag and Smets 2008a, b; Baltag et al. 2019a; van Benthem 2007, 2011), as well as in the "dynamic interactive epistemology" approach developed by game-theorists (Board 2004). These are in fact adaptations to doxastic modeling of the older setting of Lewis spheres, with its equivalent description in terms of a comparative similarity relation (Lewis 2000). In these models, the elements of M are taken to be *possible worlds*, or possible 'states' of the world, and the structure is purely qualitative, given either in terms of a nested, exhaustive family of spheres, or in terms of a total preorder on worlds. Sometimes an additional converse well-foundedness condition, or a weaker 'Limit Condition', is imposed to ensure the existence of maximal elements $Max(M) \neq \emptyset$ (or equivalently, the existence of a smallest sphere). As seen below, this simplifies the definition of (conditional) belief, as the doxastic analogue of Lewis conditionals. But as noted by Lewis (2000), such additional assumptions are not really needed, since a satisfactory notion of conditional (or conditional belief) can still be defined in non-converse-wellfounded models. Hence, we make no such additional assumptions here.

Our models are just a special case of plausibility models, adapted to a probabilistic setting: the possible worlds come as probability distributions, while the plausibility preorder and the Grove spheres are quantitatively defined from a plausibility ranking map. But the mechanism for forming beliefs $B(P)$ and conditional beliefs $B(P|Q)$

[12] Take a coin, for which there is no reason to suspect an in-built bias. Initially, before receiving any other information, the set of possible distributions was [0, 1] (if we represent each distribution by the probability it assigns to *Heads*), and the most plausible distribution was the fair one μ_{eq} (assigning probability 0.5 to *Heads*). But in the meantime, one piece of new higher-order information was received, namely that the coin is *not* perfectly fair (due to some small manufacture accidents). *Now*, μ_{eq} is excluded as impossible, so the set of possibilities is $M = [0, 1]\backslash\{0.5\}$, but nevertheless, there is still no reason to suspect any systematic bias. So, the distributions that are closer to μ_{eq} have higher plausibility, and their plausibility decreases as we move further away from it. The *only* way to extend this plausibility in a continuous way to the closure $\overline{M} = [0, 1]$ is to continue to assign maximal plausibility to μ_{eq}. This merely technical constraint makes also conceptual sense, if we think counterfactually: *if* the received information happened to be wrong, then we'd revert to considering μ_{eq} as the most plausible distribution. A priori, this impossible distribution is still inherently the more plausible.

in our probabilistic plausibility models will be exactly the same as in the general (non-converse-wellfounded, non-probabilistic) plausibility models.

Connections to inference processes Our probabilistic plausibility models can also be seen as a generalization and refinement of Paris' *inference processes* (Paris 1994; Paris and Rad 2008; Paris and Vencovska 1997). Roughly speaking, an inference process is a map Bel assigning to each set $M \subseteq M_O$ of distributions some "believed" distribution $Bel(M) \in M$. The definition in Paris (1994) actually restricts the domain of Bel to a subclass of $\mathcal{P}(M_O)$ (namely the ones definable by a set of linear inequalities),[13] but our more general setting extends this to all sets of distributions. A good look at Paris' examples of interesting inference processes shows that all of them define the salient distribution *Bel(M)* by *maximizing (or minimizing) over M a certain continuous quantity* (entropy, distance from centre of mass, distance from barycentre, etc). Our approach makes explicit this method of generating inference processes, in the form of the plausibility map, and recognizes it as just a special case of the standard method of belief formation in Logic and Belief Revision Theory. Generalizing to arbitrary sets of distributions also forces us to give up on the insistence for only one most preferred distribution,[14] or even a set of most preferred distributions. Following Lewis' approach (Lewis 2000) (as later adapted to non-converse-wellfounded plausibility models), one can still define beliefs as we did above, in terms of propositions that hold on all distributions that are plausible enough. Indeed, this seems the most natural generalization of maximization-based inference processes to arbitrary sets.

In closed models (and more generally in models in which plausibility map attains its maximum value 1) the definition of belief can be simplified, yielding the maximization-based notion of belief that is standard in both inference processes and Belief Revision Theory (in terms of maximizing plausibility rank). In such cases, *belief amounts to truth in all the 'most plausible' distributions* (the ones with plausibility rank 1):

Proposition 4 *If* $\mathbf{M} = (M, pla)$ *is a closed model, then there exists some* $\mu \in M$ *with* $pla(\mu) = 1$*; i.e. we have* $Max(M) \neq \emptyset$.[15] *Moreover, whenever* $\mathbf{M} = (M, pla)$ *is any model with* $Max(M) \neq \emptyset$ *(and hence in particular, whenever M is closed) and* $P \subseteq M_O$ *is any proposition, we have that: P is believed iff* all distributions in M with plausibility 1 satisfy *P; i.e.*

$$B(P) \text{ holds in } \mathbf{M} \quad \text{iff} \quad Max(M) \subseteq P.$$

Proof For the first part, let $M \subseteq M_O$ be closed. Since *pla* is a continuous function, we can use Propositions 1, 2(1) and 3 , to conclude that *pla* attains its supremum on M, hence $M^1 = Max(M) \neq \emptyset$.

[13] In fact, there are other differences: Paris' approach is syntactic, so the linear inequalities involve probabilities of sentences in a given language.

[14] The existence of maximizers in Paris (1994) is ensured by the fact that the sets defined by linear inequalities are closed, while the quantity to be maximized is continuous. The uniqueness of the maximizer is ensured there by the fact that these sets are convex, while the relevant quantity is concave (or convex, in the case of minimization).

[15] Recall that $Max(M) = M^1 = \{\nu \in M \mid pla(\mu) = 1\} = \{\nu \in M \mid \nu \geq^{\mathbf{M}} \nu' \text{ for all } \nu' \in M\}$, if non-empty, is the smallest Grove sphere.

For the second part, assume only that $Max(M) \neq \emptyset$. To prove the left-to-right direction in the displayed equivalence, suppose that *B(P)* holds; then by definition, there exists $\delta \leq 1$ such that $\emptyset \neq M^\delta \subseteq P$. But $\delta \geq 1$ implies $M^1 \subseteq M^\delta$, hence by transitivity of inclusion we conclude that $Max(M) = M^1 \subseteq P$, as desired.

For the converse, suppose that we have $Max(M) = M^1 \subseteq P$. Since $Max(M) \neq \emptyset$, take any $\mu \in Max(M) = M^1$. Then we have $\{\nu \in M \mid pla(\nu) \geq pla(\mu) = 1\} = M^1$ (since *pla* cannot take values larger than 1), hence $\{\nu \in M \mid pla(\nu) \geq pla(\mu)\} \subseteq P$, i.e. *B(P)* holds in $\mathbf{M}$. □

Some canonical plausibility maps and plausibility-generating functions

Here are some specific examples:

1. *Entropy-based* plausibility maps: The most direct implementation of the Principle of Indifference is to take as our generating plausibility measure the *Shannon entropy* $Ent : M_O \to [0, \infty)$, given by putting

$$Ent(\mu) := -\sum_{o \in O,\, \mu(o) \neq 0} \mu(o) \log(\mu(o))$$

It is convenient to assume that the logarithms are taken in base n (where recall that $n = |O|$ is the number of outcomes in O). This measure generates a plausibility model $\mathbf{M} = (M, Ent^M)$ on every non-empty set $M \subseteq M_O$ of distributions *with positive entropy*. The generated probabilistic plausibility map Ent^M is obtained by renormalizing entropy wrt M, as described in the generic example above: for $\mu \in \overline{M}$, put $Ent^M(\mu) := \frac{Ent(\mu)}{Ent(M)}$, where $Ent(M) := sup\{Ent(\nu) \mid \nu \in M\}$. So the most plausible distribution will be the one with highest Shannon entropy, i.e. the *most uninformative* one.[16] More generally, less informative distributions will be more plausible than more informative ones. Note also that, when using logarithms in base $n = |O|$, we have $Ent(M_O) = Ent(\mu^{eq}) = \sum_{1 \leq i \leq n} -\frac{1}{n} log_n \frac{1}{n} = 1$ (where μ^{eq} is the distribution that gives equal probability $\frac{1}{n}$ to every outcome), hence $Ent^{M_O} = Ent$.

One of the "defects" of entropy *Ent* as a plausibility-generating measure is that it may take value zero, so it is not fully positive. This means there exist non-empty sets of distributions M, for which *(M, Ent)* is technically speaking *not* a plausibility model (since $Ent(\mu) = 0$ for some $\mu \in M$): indeed, the set M_O of all distributions is such a counterexample! But recall that only the plausibility (pre-)order $\leq^{\mathbf{M}}$ is of relevance when forming beliefs. So we can take instead any positive continuous function that induces the same order. One simple way to do this is to add to entropy some fixed positive number, say 1. In this way we obtain a *fully positive version of entropy measure* $Ent^+ : M_O \to (0, \infty)$, given by putting

$$Ent^+(\mu) := 1 + Ent(\mu) = 1 - \sum_{o \in O,\, \mu(o) \neq 0} \mu(o) \log(\mu(o)).$$

[16] This mechanism for prior belief-formation matches the so-called entropic inference process (Paris 1994).

Using Ent^+ as our plausibility-generating measure, we generate a plausibility model (M, Ent^{+M}) on *every* non-empty set $M \subseteq M_O$, whose plausibility map is once again obtained by renormalizing Ent^+ to M. Moreover, Ent^{+M} agrees with Ent^M on the ranking order between any two distributions, so it induces the same plausibility ranking order as the one given by entropy. As a consequence, for every plausibility model (M, Ent^M), all beliefs and conditional beliefs (as well as knowledge) are the same as in the model (M, Ent^{+M}).
Philosophically speaking, taking either *Ent* or Ent^+ as one's plausibility measure amounts intuitively to the adoption of the Principle of Indifference at the level of the possible outcomes.

2. *Cautious* plausibility: The most 'cautious' choice of plausibility is assigning *equal plausibility* to all possible distributions, e.g. taking

$$C(\mu) := 1 \qquad \text{for all } \mu \in M_O.$$

Obviously, this is a fully positive plausibility measure, so it induces a plausibility model on every non-empty set $M \subset M_O$ (with the generated plausibility map given by the restriction of C to $\overline{M}$).
Cautious plausibility can be thought of as yet another application of the Principle of Indifference at a higher level (that of all possible distributions): since a priori there is no reason to prefer a distribution to another, the prior plausibility assigns equal rank to all of them. With this cautious choice, the prior beliefs do not go beyond what is known: *the agent only believes what she knows*. (But as we'll see, this is no longer the case after more information is received, e.g. via sampling evidence from the unknown distribution.)

3. *Typicality-based* plausibility maps: The so-called Limiting Centre of Mass of a set of distributions $M \subseteq M_O$ (also called Centre of Mass Infinity) is the output of a probabilistic inference process (Paris 1994), that involves maximizing the quantity $\sum_{o \in O(M)} \log(\mu(o))$, where we fixed a set $M \subseteq M_O$ and used the notation $O(M) = \{o \in O \mid \exists \mu \in M \text{ with } \mu(o) > 0\}$. Whenever it exists and is unique (as e.g. in the case of closed and convex sets M), the maximizer of this function over a set of distributions $M \subseteq M_O$ gives a form of 'averaging' over M. So, in general, distributions for which this quantity has a higher value are closer to the average of M.
Unfortunately, the function $\sum_{o \in O(M)} \log(\mu(o))$ takes *no positive values*, since its range is $[-\infty, 0)$. But once again, only the induced ranking order is of relevance when forming beliefs, so we can apply any continuous transformation from $[-\infty, 0)$ to $[0, 1)$ (e.g. $x \mapsto 2^x$), to obtain a plausibility measure

$$CM_\infty(\mu) := 2^{\sum_{o \in O(M)} \log(\mu(o))} = \prod_{o \in O(M)} \mu(o),$$

where here we assumed that the logarithm is taken in binary base. Assume now that $M \subseteq M_O$ is a non-empty set with the property that for every outcome $o \in O$, we have either $\mu(o) = 0$ for all $\mu \in M$ or else $\mu(o) > 0$ for all $\mu \in M$. Then the measure CM_∞ generates a probabilistic plausibility map on M, obtainable once

again by renormalization to M.

If we instead apply first a slightly different transformation ($x \mapsto 1 + 2^x$), we can go further and convert CM_∞ into a *fully positive* plausibility measure CM^+_∞. This helps avoid any restrictions on M: as long as $M \neq \emptyset$, CM_∞ generates a probabilistic plausibility map CM^{+M}_∞ on M, that induces the same preorder $\leq^{\mathbf{M}}$ as the original function $\sum_{o \in O(M)} \log(\mu(o))$. Hence, (M, CM^{+M}_∞) is a plausibility model for every non-empty M, and its ranking order, beliefs, conditional beliefs etc, agree with the one of (M, CM^M_∞), whenever the second is a plausibility model.

Taking CM^M_∞ or CM^{+M}_∞ as one's plausibility ranking amounts intuitively to the adoption of a Principle of 'Averageness' or Typicality. Indeed, the probability distributions in M that have a higher CM^{+M}_∞-plausibility will be those that are "more typical", more 'normal' or representative for M; while the most plausible ones are the "most typical".

Another typicality-based plausibility map is related to the *barycentre inference process* (Paris 1994): this involves *minimizing* the function

$$\mu \mapsto \sup\{d(\mu, \nu) \mid \nu \in M\}$$

If it exists and is unique, the minimizer of this function over M is called the *barycentre* of the set M, and it gives another notion of averageness or representativeness. It chooses the distribution μ that minimizes the worst error that could be made (when one wrongly takes μ to be the true probability). To convert this into a maximization problem, we can apply the transformation 2^{-x}, obtaining the (fully positive) *barycentric plausibility measure* $BM : M \to (0, 1]$, for any non-empty set $M \subseteq M_O$ and arbitrary distribution $\mu \in \overline{M}$:

$$B^M(\mu) := 2^{-sup\{d(\mu,\nu) \mid \nu \in M\}}.$$

Using again renormalization, this generates a probabilistic plausibility map on M, that will assign higher plausibility to distributions that are closer to M's barycenter.

4. *Evidence-based* plausibility: Given an observed event $e \in \mathcal{E}$, we may prefer distributions that maximize the probability of e. This corresponds to taking as our plausibility measure the function F_e from Lemma 1, given by $F_e(\mu) = \mu(e)$. This gives higher ranking to distributions that assign higher probability to the event e. When renormalized to any non-empty set $M \subseteq M_O$ with the property that $\mu(e) > 0$ for all $\mu \in M$, it induces a plausibility model (M, F^M_e), given by $F^M_e(\mu) := \frac{\mu(e)}{sup\{\nu(e) \mid \nu \in M\}}$.
5. *Centered plausibility*: Given a salient distribution μ (that is considered as the most plausible), one may adopt a plausibility map given by a "normal" curve centered at μ. This means that distributions that are closer to μ are considered more plausible than the ones that are farther: $pla(\nu) \geq pla(\nu')$ iff $d(\nu, \mu) \leq d(\nu', \mu)$. One example of a fully positive plausibility measure that induces this ranking order is $C_\mu : M_O \to (0, 1]$, given by putting $C_\mu(\nu) := 2^{-d(\nu,\mu)}$.

6. *Plausibility based on second-order probability*: Let $M \subseteq M_O$ be a discrete[17] set of distributions, and let $P : M \to [0, 1]$ be any second-order probability mass function (cf. Gaifman and Snir 1982; Gaifman 2016), that is required to satisfy $P(\mu) > 0$ for all $\mu \in M$ and $\sum_{\mu \in M} P(\mu) = 1$. Then this function can be extended to a continuous function $P : \overline{M} \to [0, 1]$, by putting $P(\mu) := 0$ for all limit points $\mu \in \overline{M} \backslash M$. The fact that this extension is continuous follows from the assumption that $\sum_{\mu \in M} P(\mu) = 1$, which implies that $lim_{n \to \infty} P(\mu_n) = 0$ for any infinite sequence of distinct points $\mu_n \in M$. By taking this extended function P as our plausibility measure, we generate a plausibility model (M, P^M), by renormalizing as above: $P^M(\mu) := \frac{P(\mu)}{sup\{P(\mu) \mid \mu \in M\}} = \frac{P(\mu)}{max_M(P)}$. (Note that, in order for $\sum_{\mu \in M} P(\mu)$ to have a finite value, P must attain a maximum value $max_M(P) := max\{P(\mu) \mid \mu \in M\}$ on M.)
   However, note that the beliefs based on the plausibility ranking P^M will *not* necessarily match the Lockean beliefs based on the second-order probability P. Only the distributions $\mu \in Max(M)$, having $P^M(\mu) = 1$, or equivalently $P(\mu) = max_M(P)$, are relevant for the agent's plausibilistic beliefs: she will believe that the true distribution is one of the ones in $Max(M)$. This will hold *even in the case that* $\sum_{\mu \in Max(M)} P(\mu) < \frac{1}{2}$; while an agent using P as her second-order probability will have in this case *precisely the opposite belief*: she believes that the true distribution is in $M \backslash Max(M)$, since this is more likely to be the case. This points yet again to the *fundamental difference* between the interpretation of a function as a plausibility map versus its meaning as a probability function. Plausibility ranks do not obey the Kolmogorov additivity axiom, but instead higher plausibility ranks simply dominate lower ones.

Example 1 (*continued*). In the Coin example, we initially have no information about the coin, the set of possible coin biases will be the set M_O of *all* probability mass functions on $O = \{H, T\}$. Suppose that we have background information that the extremal distributions (μ_0 with $\mu_0(H) = 0$, and μ_1 with $\mu_1(H) = 1$) are impossible. Then the set of possibilities is given by $M := M_O \backslash \{\mu_0, \mu_1\}$. We can choose the *entropy Ent* as our plausibility map, as this can be justified here in terms of symmetry: the faces of a coin (or a dice) are symmetric, so there is no reason to prefer one outcome over another. Then (M, Ent^+) is a plausibility model, where the highest plausibility will be given to the distribution with the highest entropy: the fair-coin distribution μ^{eq}, assigning $\mu^{eq}(H) = \mu(T) = \frac{1}{2}$ (since for every $\nu \neq \mu^{eq}$ we have $Ent(\nu) < Ent(\mu^{eq})$). So entropic plausibility starts with an initial belief in the fairness of the coin (and more generally it assigns a higher ranking to a distribution that corresponds to a more well-balanced coin). Note that *entropic plausibility induces the same ranking order on this model as the centered plausibility* $C_{\mu^{eq}}$ (centered at the fair-coin distribution μ^{eq}).

If, however, we cannot exclude *any* distribution (not even the extremal ones), then the set of possibilities is the whole M_O, and Ent will no longer give us a plausibility model. Still, we can choose instead the positive version of entropic plausibility Ent^+, which makes (M_O, Ent^+) into a plausibility model, while maintaining the same initial

[17] A set is discrete if it consists only of isolated points. Discrete subsets of M_O are either finite or countable.

belief in the coin's fairness (and the same preference for more well-balanced coins). Note again that Ent^+ still induces the same ranking order on this model as the centered plausibility $C_{\mu^{eq}}$.

Example 2 (*continued*). In the Urn example, we initially have no other information besides the three colors, so the set of possibilities is the set M_O of all distributions over $O = \{R, B, G\}$. Since there is no reason to prefer any one distribution over any other (and no considerations of symmetry are relevant, since we cannot see inside the urn to somehow assess whether there is a rough balance between the quantities of marbles of different colors), the most natural prior ranking seems to be in this case the *cautious plausibility* C: each possible distribution is assigned an equal plausibility of 1. In the plausibility model (M_O, C), the agent has no other rational beliefs at this stage, beyond what she knows.[18]

Proposition 5 *Let* $\mathbf{M} = (M, pla)$ *be any model, then Knowledge and belief satisfy the following properties:*

1. *Knowledge is truthful: if K(P) holds, then P holds at all possible distributions (i.e.* $M \subseteq P$*);*
2. *Tautologies are known:* $K(M_O)$ *holds;*
3. *Knowledge implies belief: if K(P) holds then B(P) holds.*
4. *Belief is consistent:* $B(\emptyset)$ *never holds;*
5. *Knowledge and belief are closed under entailment: if* $P \subseteq Q$*, then K(P) implies K(Q), and similarly B(P) implies B(Q);*
6. *Knowledge and belief are (finitely) conjunctive: if* $K(P_i)$ *holds for all* $1 \leq i \leq n$*, then* $K(\bigcap_{i=1}^n P_i)$ *holds; similarly, if* $B(P_i)$ *holds for all* $1 \leq i \leq n$*, then* $B(\bigcap_{i=1}^n P_i)$ *holds;*
7. *Any finite number of beliefs are mutually consistent: if* $B(P_i)$ *holds for all* $1 \leq i \leq n$*, then* $\bigcap_{i=1}^n P_i \neq \emptyset$*.*

Proof Properties 1,2,3,4 follow immediately from the definitions of knowledge and belief. Property 5 for knowledge follows directly from property 1. For property 5 for belief: *B(P)* gives the existence of some $\delta > 0$ with $\emptyset \neq M^\delta \subseteq P$, which together with $P \subseteq Q$ gives us $\emptyset \neq M^\delta \subseteq Q$, hence $B(Q)$ holds. Property 6 for knowledge follows from property 1, via the sequence of implications: if $K(P_i)$ holds for all $1 \leq i \leq n$, then $M \subseteq P_i$ for all $1 \leq i \leq n$, so $M \subseteq \bigcap_{i=1}^n P_i$, hence $K(\bigcap_{i=1}^n P_i)$ holds. Property 6 for belief: suppose that $B(P_i)$ holds for all $1 \leq i \leq n$; so, for every $1 \leq i \leq n$, there exists some $\delta_i > 0$ s.t. $\emptyset \neq M^{\delta_i} \subseteq P_i$. Take $\delta = min\{\delta_i \mid 1 \leq i \leq n\}$. Then we have $\delta > 0$, $M^\delta \neq \emptyset$, and $M^\delta \subseteq \bigcap_{i=1}^n M^{\delta_i} \subseteq \bigcap_{i=1}^n P_i$, hence $B(\bigcap_{i=1}^n P_i)$ holds. Property 7 follows immediately from properties 6 and 4. □

Finally, one should note that belief in closed models (or more generally, any model having most plausible distributions) is better behaved, having stronger consistency and conjunctivity properties, than in arbitrary models:

[18] Similarly to Example 1, we can also consider here the case in which we are given the background information that all three possible colors actually occur, so that the extremal distributions are impossible, i.e. we have $\mu(R), \mu(B), \mu(G) \neq 0$.

Proposition 6 *Let* $\mathbf{M} = (M, pla)$ *be any model with* $Max(M) \neq \emptyset$. *Then we have the following:*

- beliefs are closed under arbitrary conjunctions*: if* $\{P_i \mid i \in I\}$ *is a family of propositions such that* $B(P_i)$ *holds for all* $i \in I$, *then* $B(\bigcap_{i\in I} P_i)$ *also holds;*
- *beliefs are* globally consistent*:* $\bigcap\{P \subseteq M_O \mid B(P)$ *holds in* $\mathbf{M}\} \neq \emptyset$.

In particular, these properties hold in closed *models.*

Proof For the first item, suppose that $B(P_i)$ holds for all $i \in I$. Then by the second part of Proposition 4, we have $Max(M) \subseteq P_i$ for all i, and hence $Max(M) \subseteq \bigcap_{i\in I} P_i$, hence $B(\cap_{i\in I} P_i)$ (again by Proposition 4).

For the second item, we apply the first item to the family $\{P \subseteq M_O \mid B(P)$ holds in $\mathbf{M}\}$ to infer that we have $B(\{P \subseteq M_O \mid B(P)$ holds in $\mathbf{M}\})$, then apply Proposition 5.3 to obtain the desired conclusion. □

The following example shows that the above properties do *not* necessarily hold in arbitrary probabilistic plausibility models!

Counterexample: Suppose that, in the Coin Example, our agent learns from the manufacturer only one piece of information: the coin is *not* completely fair, due to very small, accidental imperfections (rather than any intentional bias). What is a rational agent, who forms entropy-based beliefs, supposed to believe? Smaller imperfections seem to be more plausible than larger ones: hence, any bias closer to $\frac{1}{2}$ is more plausible than one that is farther. On the other hand, the agent knows for sure that the coin is not fair. Our agent has acquired omega-inconsistent beliefs, which nevertheless seem rational, given her information.

To formalize this counterexample, take $O = \{H, T\}$ as in the Coin Example, and take the model (M, Ent^+) with $M = M_O \backslash \{\mu^{eq}\}$, where $\mu^{eq}(H) = \mu(T) = \frac{1}{2}$ is the fair-coin distribution and Ent^+ is the positive version of entropic plausibility. Recall that Ent^+ yields on the same ranking order on M_O as the centered plausibility $C_{\mu^{eq}}$: distributions that are closer to μ^{eq} are more plausible than the ones that are farther. For each $n \geq 2$, take $P_n := \{\mu \in M \mid \mu(H) \in (\frac{1}{2} - \frac{1}{n}, \frac{1}{2} + \frac{1}{n})\}$. Then $B(P_n)$ holds for all $n \geq 2$ (since every distribution close enough to μ^{eq} is in P_n), but $\bigcap_{n\geq 2} P_n = \emptyset$ (since $\mu^{eq} \notin M$), hence beliefs are globally inconsistent; moreover, $B(\bigcap_{n\geq 2} P_n)$ does not hold (since $B(\emptyset)$ is false, by Proposition 5.3), hence *beliefs are not necessarily closed under countable conjunctions*.

This counterexample shows that plausibility-based beliefs in non-closed models may be subject to a kind of Infinite Lottery Paradox: though believing, for each $n \geq 2$, that the coin's bias is in $(\frac{1}{2} - \frac{1}{n}, \frac{1}{2} + \frac{1}{n})\backslash\{\frac{1}{2}\}$, our agent does *not* believe that the bias is in (empty) intersection of all these sets. So beliefs in non-closed models may exhibit a type of 'omega-inconsistency': though each belief is consistent, and any finitely many beliefs are mutually consistent, the family of all beliefs may still be inconsistent, when taken as a whole!

We think this is a small price to pay for being able to form beliefs when given arbitrary information $M \subseteq M_O$. Situations such as in the above counterexample can occur in practice, whenever partial information is obtained, say by communication. Still, readers who consider global doxastic consistency to be an inherent feature of rationality are welcome to restrict our framework to models in which the plausibility

map attains a maximum value. Full infinitary conjunctivity and global consistency of beliefs can be regained in this way, without any other loss, except for generality.

4 Conditioning and belief dynamics

One of the main motivations for developing the setting that we investigate here is to capture the *process of learning a distribution* as a form of iterated belief revision, that results from receiving new information. But, as already explained, the two components of our probabilistic plausibility models $\mathbf{M} = (M, pla)$ capture *two different types of information* about the unknown distribution μ: the set M represents the agent's hard higher-level information about μ (her 'knowledge', given by the proposition $M \subseteq M_O$); while the plausibility map $pla : M_O \to [0, 1]$ represents the agent's soft information about μ (typically obtained by sampling or other observational events), her "beliefs" given by the ranking order. Each of these two forms of information is subject to its own type of revision, captured by its own form of conditioning or update: (1) *conditioning on a new proposition* $Q \subseteq M_O$, resulting in an eliminative update with the hard information Q, by which some distributions are eliminated, while the plausibility ranking stays the same; (2) *conditioning on a new observational event* $e \in \mathcal{E}$ (resulting in an upgrade of the plausibility map, by which distributions assigning a higher probability to e get a boost ranking, while the set M typically stays the same (except possibly for the elimination of those extreme distributions that assigned zero probability to e).

Definition 4 [*Two forms of conditioning and updating*] Given a plausibility model $\mathbf{M} = (M, pla)$, a proposition $P \subseteq M_O$ is said to be *compatible with* $\mathbf{M}$ if the intersection $M \cap P \neq \emptyset$ is non-empty. Similarly, an event $e \in \mathcal{E}$ is said to be *compatible with* $\mathbf{M}$ if there exist distributions $\mu \in M$ with $\mu(e) \neq 0$, i.e. the set $M_e := \{\mu \in M \mid \mu(e) \neq 0\}$ is non-empty.

Let $Prop_{\mathbf{M}}$ be the family of all propositions compatible with $\mathbf{M}$, and let $\mathcal{E}_{\mathbf{M}}$ be the family of all events compatible with $\mathbf{M}$. We can define two binary operations $pla(.|.) : \overline{M} \times Prop_{\mathbf{M}} \to [0, 1]$ (*conditioning on a proposition*) and $pla(.|.) : \overline{M} \times \mathcal{E}_{\mathbf{M}} \to [0, 1]$ (*conditioning on an event*), by putting

$$pla(\mu|P) := \frac{pla(\mu)}{pla(M \cap P)} = \frac{pla(\mu)}{sup\{pla(\nu) \mid \nu \in M \cap P\}},$$

$$pla(\mu|e) := \frac{pla(\mu, e)}{pla(e)} = \frac{pla(\mu, e)}{pla(M, e)} = \frac{pla(\mu) \cdot \mu(e)}{sup\{pla(\nu) \cdot \nu(e) \mid \nu \in M\}}.$$

The two types of conditioning give rise to two forms of dynamic operations on models, corresponding to two distinct varieties of learning: updating the plausibility model $\mathbf{M} = (M, pla)$ with a compatible proposition $P \in Prop_{\mathbf{M}}$ yields the *P-updated model* $\mathbf{M}_P = (M_P, pla_P)$, given by $M_P := M \cap P$ and $pla_P(\mu) := pla(\mu \mid P)$ for $\mu \in \overline{M \cap P}$; while updating the same model with a compatible event $e \in \mathcal{E}_{\mathbf{M}}$ yields the *e-updated model* $\mathbf{M}_e = (M_e, pla_e)$, given by $M_e := \{\mu \in M : \mu(e) \neq 0\}$ and $pla_e(\mu) := pla(\mu \mid e)$ for $\mu \in \overline{M_e}$.

The first type of conditioning can be recognized as a *plausibilistic analogue of the Kolmogorov definition of conditional probability*, that fits well with *propositional learning*. Note that the *P-conditional plausibility order* $\leq_P$ in the model $\mathbf{M}_P$, given by

$$\mu \leq_P \mu' \text{ iff } pla_P(\mu) \leq pla_P(\mu') \text{ iff } pla(\mu) \leq pla(\mu') \text{ iff } \mu \leq \mu',$$

is *the same* as the initial plausibility order $\leq$, except that it is restricted to $M \cap P$ (since the renormalizing denominator $pla(M \cap P)$ in the definition of pla_P doesn't make a difference for the order). Indeed, the propositional update (generated by receiving new "hard" higher-order information P) shrinks the space of possible distributions M by eliminating certain possibilities, while leaving the plausibility map "essentially the same" (modulo the renormalizing factor). This shows that our propositional update falls well within the scope of traditional Belief Revision Theory, representing a special case of AGM conditioning.

On the other hand, the second type of conditioning can be seen as a *plausibilistic analogue of Bayes' conditioning formula* (where in both cases, the operation *sup* of taking supremum plays the role usually played by addition $\sum$), and thus captures a notion of *learning through sampling*. The event conditioning rule weights the plausibility of each distribution with how well it predicts the observed sampling event e. Note that *e-conditional plausibility order* $\leq_e$ in the model M_e is given by

$$\mu \leq_e \mu' \text{ iff } pla_e(\mu) \leq pla_e(\mu') \text{ iff } pla(\mu) \cdot \mu(e) \leq pla(\mu') \cdot \mu'(e).$$

Indeed, the event update is generated by receiving "soft" information (obtained by sampling), and it naturally resembles soft doxastic 'upgrades' (rather than updates) from Dynamic Epistemic Logic (Baltag and Renne 2016; van Benthem 2011; Baltag and Smets 2008b): it leaves the set of possibilities M "essentially the same" (since it does not necessarily eliminate any distribution, except for the extremal ones, assigning probability 0 to e, if there any in M), but rather only changes the plausibility over them. Distributions that better fit the sampling evidence are only 'promoted' in plausibility, while the others are demoted (but not eliminating, except for the extremal ones).

The next result confirms that our updates are well-defined operations on plausibility models:

Proposition 7 *Let* $\mathbf{M} = (M, pla)$ *be a plausibility model,* $P \in Prop_{\mathbf{M}}$ *be a compatible proposition, and* $e \in \mathcal{E}_{\mathbf{M}}$ *be a compatible event. Then* $\mathbf{M}_P = (M_P, pla_P)$ *and* $\mathbf{M}_e = (M_e, pla_e)$*, as defined above, are plausibility models.*

Proof For propositional updates, the compatibility of P with $\mathbf{M}$ implies that the domain of the P-updated model is non-empty: $M_P = M \cap P \neq \emptyset$. Similarly, the compatibility of the event e with $\mathbf{M}$ implies that $M_e \neq \emptyset$. It is easy to check that the function $pla_P : \overline{M \cap P} \to [0, \infty)$, given by $pla_P(\mu) = \frac{pla(\mu)}{pla(M \cap P)}$, takes indeed values in $[0, 1]$ (since $0 < pla(\mu) \leq max\{pla(\nu) \mid \nu \in \overline{M \cap P}\} = sup\{pla(\nu) \mid \nu \in M \cap P\} = pla(M \cap P)$ for $\mu \in \overline{M \cap P}$); moreover, $pla_P(\mu) > 0$ for all $\mu \in M_P = M \cap P$ (since the denumerator $pla(\mu) > 0$ for all $\mu \in M$); and finally $sup\{pla^P(\nu) \mid \nu \in$

$M \cap P\} = sup\{\frac{pla(\nu)}{pla(M\cap P)} \mid \nu \in M \cap P\} = \frac{sup\{pla(\nu) \mid \nu\in M\cap P\}}{pla(M\cap P)} = 1$ (again using the fact that $sup\{pla(\nu) \mid \nu \in M \cap P\} = pla(M \cap P)$).

Similarly, the definition of M_e ensures that the function $pla_e(\mu) = pla(\mu \mid e) = \frac{pla(\mu)\cdot\mu(e)}{sup\{pla(\nu)\cdot\nu(e) \mid \nu\in M_e\}}$ takes only positive values on M_e, and that its supremum is 1 on M_e. To show that pla_e is continuous, we put together the definition of conditional plausibility, the fact that $pla_e = \frac{pla\cdot F_e}{k}$ (where F_e is the function introduced in Lemma 1 and $k = sup\{pla(\nu) \cdot \nu(e) \mid \nu \in M_e\}$ is a non-zero constant), the continuity of pla (by definition) and of F_e (by Lemma 1), and use the closure of continuous functions under products and division by non-zero constants. □

This fact allows us to *iterate* and even *interleave* the two forms of updating. For simplicity, we only do it for events and propositions that *fit the true distribution* (since this automatically ensures their mutual compatibility):

Definition 5 (*Iterated updating*) Given a plausibility model $\mathbf{M} = (M, pla)$, and let $\mu \in M$ be the 'true' distribution, we can define the *iterated update* $\mathbf{M}_\sigma$, for every finite sequence $\sigma = (\sigma_1, \ldots, \sigma_n) \in (Prop \cup \mathcal{E})^*$ consisting of true propositions ($\sigma_i \in Prop$ with $\mu \in \sigma_i$) or truly observable events ($\sigma_i \in \mathcal{E}$ with $\mu(\sigma_i) \neq 0$). The definition is by recursion on the length of the σ, by putting:

$$\begin{aligned}
\mathbf{M}_\lambda &:= \mathbf{M}, \quad \text{for the empty sequence } \lambda = (),\\
\mathbf{M}_{\sigma,e} &:= (\mathbf{M}_\sigma)_e, \quad \text{for observable } e \in \mathcal{E} \text{ (with } \mu(e) \neq 0),\\
\mathbf{M}_{\sigma,P} &:= (\mathbf{M}_\sigma)_P, \quad \text{for truthful } P \in Prop \text{ (with } \mu \in P).
\end{aligned}$$

The next three results ensure that updating satisfies some standard rationality constraints: Proposition 8 guarantees that the result of repeated conditionalisation is independent of the order of application; Proposition 9 says that the result of conditioning is independent of whether it is done successively (conditioning on each independent observation, one after the other) or in one global step (conditioning on the whole sequence of independent observations, as one big single event); while Proposition 10 shows that, when conditioning with a sequence of observations, the result is independent of the temporal order of the observations. These last three facts are important as they ensure that the agent's posterior beliefs depend only on the *evidence* that is observed (and the prior plausibility model), not on the temporal or logical order in which this evidence is observed or processed.

Proposition 8 *The order of applying (iterated) conditionalization is irrelevant: if $\sigma, \sigma' \in (Prop \cup \mathcal{E})^*$ are sequences of equal length m of propositions and/or events, s.t. σ' is obtained by permuting the components of σ (i.e. there exists some bijection $g : \{1, \ldots, m\} \to \{1, \ldots, m\}$ s.t. $\sigma'_i = \sigma_{g(i)}$ for all i), then we have*

$$\mathbf{M}_\sigma = \mathbf{M}_{\sigma'}$$

Proof It is enough to show that we can commute the order of any two basic updates, since then the desired conclusion follows by induction. So we only need to check that we have $\mathbf{M}_{P,e} = \mathbf{M}_{e,P}$, $\mathbf{M}_{P,Q} = \mathbf{M}_{Q,P}$ and $\mathbf{M}_{e,e'} = \mathbf{M}_{e',e}$. This is an easy but

tedious verification, so we only sketch here the last case: for the underlying set we have $M_{e,e'} = (M_e)_{e'} = \{\nu \in M_e \mid \nu(e') \neq 0\} = \{\nu \in M \mid \nu(e) \neq 0, \nu(e') \neq 0\}$, which immediately gives us $M_{e,e'} = M_{e',e}$; for the plausibility map, we have $pla_{e,e'}(\mu) = \frac{pla_e(\mu)\cdot\mu(e')}{sup\{pla_e(\nu)\cdot\nu(e') \mid \nu\in M_{e,e'}\}} = \frac{pla(\mu)\cdot\mu(e)\cdot\mu(e')}{sup\{pla(\nu)\cdot\nu(e)\cdot\nu(e') \mid \nu\in M_{e,e'}\}}$, which again immediately give us $pla_{e,e'} = pla_{e',e}$. □

The next proposition shows that conditioning successively on a number of independent observations is the same as conditioning on the single event consisting of the whole sequence of observations:

Proposition 9 *If* $\mathbf{M} = (M, pla)$ *is a plausibility model and events* $e, e' \in \mathcal{E}$ *are* independent wrt all distributions μ *with* $\mu \in M$*, then we have:*

$$(\mathbf{M}_e)_{e'} = \mathbf{M}_{e\cap e'}$$

As a consequence, for any event of the form $[\omega_1, \ldots, \omega_m] = \bigcap_{k=1}^m \omega_i^i$*, we have:*

$$\mathbf{M}_{\omega_1^1,\omega_2^2,\ldots,\omega_m^m} = \mathbf{M}_{[\omega_1,\ldots,\omega_m]}$$

(where recall that, for any outcome $o = \omega_i \in O$*, the event* $\omega_i^j = o^j := \{\tilde{\omega} \in \Omega \mid \tilde{\omega}_j = o = \omega_i\}$ *is the one of observing outcome* $o = \omega_i$ *at the* j*-th sampling from the unknown distribution, while* $[\omega_1, \ldots, \omega_m] = \bigcap_{i=1}^m \omega_i^i$ *is the event associated to the observational sequence* $\omega_1, \ldots, \omega_m$*).*

Proof By independence we have $\mu(e \cap e') = \mu(e) \cdot \mu(e')$, so we get $(M_e)_{e'} = \{\mu \in M \mid \mu(e) \neq 0, \mu(e') \neq 0\} = \{\mu \in M \mid \mu(e) \cdot \mu(e') \neq 0\} = \{\mu \in M \mid \mu(e \cap e') \neq 0\} = M_{e\cap e'}$. Similarly, as seen in the proof of Proposition 8, we have $pla_{e,e'}(\mu) = \frac{pla(\mu)\cdot\mu(e)\cdot\mu(e')}{sup\{pla(\nu)\cdot\nu(e)\cdot\nu(e') \mid \nu\in M_{e,e'}\}}$. Using the independence assumption, we obtain $pla_{e,e'}(\mu) = \frac{pla_{e\cap e'}(\mu)}{sup\{pla_{e\cap e'}(\nu) \mid \nu\in M_{e,e'}\}} = pla_{e\cap e'}(\mu)$.

The second claim of our Proposition follows by an easy induction from the first (given that, by the definition of μ, each event ω_j^j is independent on the event $\bigcap_{k=1}^{j-1} \omega_k^k$). □

While Proposition 8 states that the *logical* order of applying conditionalization (with both events and propositions) is irrelevant, the next result shows that the *temporal* order in which the outcomes are observed is also irrelevant:

Proposition 10 *For* $m \geq 1$*, let* g *be a permutation of the first* m *positive integers (i.e. a bijection* $g : \{1, 2, \ldots, m\} \to \{1, 2, \ldots, m\}$*). For any two events of the form* $[\omega_1, \ldots, \omega_m] = \bigcap_{k=1}^m \omega_i^i$ *and* $[\omega_{g(1)}, \ldots, , \omega_{g(m)}] = \bigcap_{i=1}^m \omega_{g(i)}^i$*, we have*

$$\mathbf{M}_{[\omega_1,\ldots,\omega_m]} = \mathbf{M}_{[\omega_{g(1)},\ldots,\omega_{g(m)}]}.$$

Proof Using the notations F_e from Lemma 1, and applying the multiplicative rule for independent events (as well as the associativity and commutativity of multiplication), we obtain: $F_{[\omega_1,\ldots,\omega_m]}(\mu) = \mu(\bigcap_{i=1}^m \omega_i^i) = \prod_{i=1}^m \mu(\omega_i^i) = \prod_{i=1}^m \mu(\omega_i) =$

$\prod_{i=1}^{m} \mu(\omega_{g(i)}) = \prod_{i=1}^{m} \mu(\omega^i_{g(i)}) = \mu(\bigcap_{i=1}^{m} \omega^i_{g(i)}) = F_{[\omega_{g(1)},\ldots,\omega_{g(m)}]}(\mu)$, for all $\mu \in M_O$. Using this, we get that $M_{[\omega_1,\ldots,\omega_m]} = \{\mu \in M \mid F_{[\omega_1,\ldots,\omega_m]}(\mu) \neq 0\} = \{\mu \in M \mid F_{[\omega_{g(1)},\ldots,\omega_{g(m)}]}(\mu) \neq 0\} = M_{[\omega_{g(1)},\ldots,\omega_{g(m)}]}$. Similarly (using also the definition of conditional plausibility), we conclude that $pla_{[\omega_1,\ldots,\omega_m]}(\mu) = \frac{pla(\mu)\cdot F_{[\omega_1,\ldots,\omega_m]}(\mu)}{sup\{pla(\nu)\cdot F_{[\omega_1,\ldots,\omega_m]}(\nu) \mid \nu \in M\}} = \frac{pla(\mu)\cdot F_{[\omega_{g(1)},\ldots,\omega_{g(m)}]}(\mu)}{sup\{pla(\nu)\cdot F_{[\omega_{g(1)},\ldots,\omega_{g(m)}]}(\nu) \mid \nu \in M\}} = pla_{[\omega_{g(1)},\ldots,\omega_{g(m)}]}(\mu)$. □

Example 1 (*continued*) Take the plausibility model (M, Ent) as before where $M := M_O \backslash \{\mu_0, \mu_1\}$ is the set of all non-extremal biases of the coin and $Ent^M : \overline{M} = M_O$ is the entropic plausibility. Since in this case $O = \{Heads, Tails\}$ has $n = 2$ outcomes, our entropy calculations will use logarithms in binary base. We have $Ent(M) = Ent(M_O) = Ent(\mu^{eq})$, where μ^{eq} is the fair distribution, so $Ent^M = Ent$. Let $e := [H, H, H] = H^1 \cap H^2 \cap H^3 \in \mathcal{E}$ be the event that "the first three tosses of the coin have landed on Heads". After observing e, no distribution is eliminated (since the only distribution incompatible with the evidence is μ_0 with $\mu(H) = 0$, which has already been excluded from the start), so $M_e = M = M_O \backslash \{\mu_0, \mu_1\}$. The new plausibility function is given by $pla_e(\mu) = \frac{pla(\mu,e)}{pla(M,e)}$, where $pla(\mu, e) = Ent(\mu) \cdot \mu(e)$. Thus the most plausible probability function will no longer be μ^{eq} and ones with a bias towards Heads will become more plausible. Let μ_1, μ_2 and μ_3 be such that $\mu_1(Heads) = 0.75$, but $\mu_2(Heads) = 0.8$ and $\mu_3(Heads) = 0.9$ then it is easy to check that $pla_e(\mu_1) < pla_e(\mu_2) > pla_e(\mu_3)$.[19] So the maximizer has $\mu(Heads) \in (0.8, 0.9)$. This is natural: the initial belief in fairness is no longer realistic; the agent now believes there is a bias towards Heads.

If however, we cannot initially exclude the extremal distributions, then Ent is not a good plausibility map, and we have to once again take its positive version to form the initial plausibility model (M_O, Ent^+). The same event $e = [H, H, H]$ will now change the model differently: it will now eliminate μ_0, yielding $M_e = M_O \backslash \{\mu_0\}$ as the new set of possibilities, while new plausibility map is given by $pla_e(\mu) := (1 + Ent(\mu)) \cdot \mu(e)$. This changes the initial belief in fairness, and distributions with a higher bias towards Heads become more plausible (though the maximizer will be slightly different than in the previous situation). Also, note that the new plausibility map still inherits from the entropic plausibility the aversion towards extremal distributions: e.g. the distribution μ_1 with $\mu_1(Heads) = 1$, though it can no longer be excluded (since $\mu_1 \in M_e$ now) and though it, in fact, matches exactly the observed frequency of Heads, will still *not* be believed (and in fact will *never* become the most plausible, after no finite sequence of observations, no matter how many times the coin falls Heads up).

Example 2 (*continued*) Take the plausibility model (M_O, C) as before where M_O is the set of all possible distributions over the set $O = \{R, G, B\}$, and C is the *cautious* plausibility. Recall that $C(\mu) = 1$ for all $\mu \in M_O$ and hence all distributions are maximizers of pla_C, so initially there are no special beliefs about the distribution.

[19] To see this notice that: $pla(\mu_1, e) = Ent(\mu_1) \times \mu_1(e) = (-0.75 \times \log_2 0.75 - 0.25 \times \log_2 0.25) \times (0.75)^3 = 0.220767$, while $pla(\mu_2, e) = Ent(\mu_2) \times \mu_2(e) = (-0.8 \times \log_2 0.8 - 0.2 \times \log_2 0.2) \times (0.8)^3 = 0.369612$, and finally $pla(\mu_3, e) = Ent(\mu_3) \times \mu_3(e) = (-0.9 \times \log_2 0.9 - 0.10 \times \log_2 0.1) \times (0.9)^3 = 0.3418977$.

The agent starts sampling marbles, noting their colour, and replacing them in the urn. Let $e := [R, R, R] = R^1 \cap R^2 \cap R^3$ be the event that "the first three sampled marbles are all Red". After observing e, all distributions μ with $\mu(R) = 0$ are eliminated, so that the new set of possibilities is $M_e = \{\mu \in M_O : \mu(R) \neq 0\}$), and the new plausibility map is given by $pla_e(\mu) = \mu(e) = \mu(R)^3$. The maximizer of this function is μ_R, given by $\mu_R(R) = 1$ and $\mu_R(G) = \mu_R(B) = 0$. So the agent now believes that there are only Red marbles in the urn: this is natural since based on her current evidence there is no reason to assume there are any Green or Blue marbles inside. If however, the next sampled marble comes up Green, then we have the event $f := e \cap G^4 = [R, R, R, G] = R^1 \cap R^2 \cap R^3 \cap G^4$. After observing this, all distributions with $\mu(G) = 0$ are also eliminated, so the new set of possibilities is $M_f = \{\mu \in M_O : \mu(R) \cdot \mu(G) \neq 0\}$. Note that the previously believed distribution μ_R has been eliminated now: not it is no longer believed, it is known now to be impossible! Furthermore, the new plausibility map is given by $pla_f(\mu) = \mu(f) = \mu(R)^3 \cdot \mu(G)$. The unique maximizer of this function is the distribution μ_{2R1G}, given by $\mu_{2R1G}(R) = \frac{2}{3}$, $\mu_{2R1G}(G) = \frac{1}{3}$ and $\mu_{2R1G}(B) = 0$. So the agent now believes that there are twice as many Red marbles than Green marbles (and no Blue marbles) in the urn. Again, this is natural, since twice as many Red marbles were observed than Green (and no Blue). One can in fact show that, when the prior is given by the cautious plausibility, the most plausible distribution after any sequence of observations will always be the one matching the observed frequencies.

The above notion of conditional plausibility gives us immediately a theory of belief revision, which can be formalized in terms of a notion of *conditional belief*. Note that this is conditionalisation on an *observable event*, corresponding to learning from observations (i.e. from sampling from the unknown distribution). On the other hand, the standard AGM setting in Belief Revision Theory and Logic (Alchourrón et al. 1985; Board 2004; Baltag and Smets 2008b; van Benthem 2011) involves revising with a *proposition* (i.e. set of distributions), rather than an event. This corresponds to learning high-level information about the unknown distribution, which allows to further shrink the range of possibilities to some subset of the prior set of possible distributions. We thus obtain *two forms of conditional beliefs*: a Bayesian-type conditioning on events, encoding 'statistical' learning; and an AGM-type of conditioning on propositions, encoding 'logical' belief revision.

Definition 6 [*Two forms of conditional belief*] Let $\mathbf{M} = (M, pla)$ be a plausibility model, and $P \subseteq M$ be a proposition. For an event $e \in \mathcal{E}$, we say that P is *believed conditional on* e in $\mathbf{M}$, and write $\mathbf{M} \models B(P|e)$, iff all e-plausible enough distributions in M are in P; i.e. for some $\mu \in M$, $\{\nu \in M \mid pla_e(\nu) \geq pla_e(\mu)\} \subseteq P$. For a proposition $Q \subseteq M$, we say that P is *believed conditional on* Q in $\mathbf{M}$, and write $\mathbf{M} \models B(P|Q)$, if and only if all plausible enough distributions in Q are in P; i.e. for some $\mu \in Q$, $\{\nu \in Q \mid pla(\nu) \geq pla(\mu)\} \subseteq P$.

It should be clear that *B(P)* is equivalent to $B(P|\Omega)$ and to $B(P|M)$, where the set Ω of all observation streams represents the *tautological event* (corresponding to "no observation") and the set M of all worlds represents the *tautological proposition* (corresponding to "no further higher-order information").

It should be equally clear that *conditional beliefs track the updated beliefs*: for every $P \subseteq M_O$, $B(P|Q)$ holds in $\mathbf{M}$ iff $B(P)$ holds in $\mathbf{M}_Q$; and similarly, $B(P|e)$ holds in $\mathbf{M}$ iff $B(P)$ holds in $\mathbf{M}_e$.[20] This allows us to generalize conditional beliefs to iterated conditioning:

Definition 7 [*General conditional belief*] Let $\mathbf{M} = (M, pla)$ be a plausibility model, and $P \subseteq M$ be a proposition. For any finite sequence $\sigma = (\sigma_1, \ldots, \sigma_n) \in (Prop \cup \mathcal{E})^*$ of propositions/and or events, we say that P is *believed conditional on* σ in $\mathbf{M}$, and write $\mathbf{M} \models B(P|\sigma)$, iff P is believed in $\mathbf{M}_\sigma$.

Conditional belief is consistent whenever the evidence is (i.e. if $e \neq \emptyset$, then $B(P|e)$ implies $P \neq \emptyset$, and similarly for $B(P|Q)$). As we'll see, beliefs conditional on events allow us to inductively learn from repeated sampling, and to ultimately converge to the true distribution. As such, they behave in a way that is somewhat similar to the usual Bayesian conditioning, used in statistical learning. In contrast, beliefs conditional on propositions will behave as a 'logical' form of belief update, satisfying all the standard axioms of Conditional Doxastic Logic (Board 2004; Baltag and Smets 2008b)(which are in fact just an equivalent formulation of the so-called AGM postulates (Alchourrón et al. 1985) from Belief Revision Theory).

As for simple belief, the definition of belief conditional on events can be simplified in closed models. In this case, *conditional belief* $B(P|e)$ *amounts to truth in all the most e-plausible distributions*:

Proposition 11 *If* $\mathbf{F} = (M, pla)$ *is a closed model and* $e \in \mathcal{E}$ *is compatible with* $\mathbf{M}$*, then there exists some* $\mu \in M_e$ *with highest e-revised plausibility in M (i.e. s.t.* $pla_e(\mu) \geq pla_e(\mu')$ *for all* $\nu' \in M_e$*). In other words, we have*

$$Max_e(M_e) \neq \emptyset,$$

where for any proposition $P \subseteq M_O$*, we put* $Max_e(P) := \{\nu \in P \mid \nu \geq_e \nu' \text{ for all } \nu' \in P\} = \{\nu \in P \mid pla_e(\nu) \geq pla_e(\nu') \text{ for all } \nu' \in P\}$.

Moreover, for any proposition $P \subseteq M_O$*, we have that P is believed conditional on e iff all most e-plausible distributions in M are in P:*

$$B(P|e) \text{ holds in } \mathbf{F} \text{ iff } Max_e(M_e) \subseteq P.$$

Proof By Proposition 7, pla_e is a plausibility function, hence it is continuous. Recall that $\mathbf{M}$ is closed and hence (by Propositions 1, 2(1) and 3) pla_e has a maximum value on M. Let $\mu \in M$ be a distribution in which this maximum value is attained, i.e. we have $pla_e(\mu) \geq pla_e(\mu')$ for all $\mu' \in M$ (and thus also for all $\mu' \in M_e \subseteq M$). Since e is compatible with $\mathbf{M}$, there exists some $\nu \in \mathbf{M}$ s.t. $\nu(e) > 0$, and hence $pla_e(\mu) \geq pla_e(\nu) = pla(\nu) \cdot \nu(e) > 0$. So we have $0 < pla_e(\mu) = pla(\mu) \cdot \mu(e)$, which implies that $\mu(e) \neq 0$, i.e. $\mu \in M_e$. This, together with the fact that $pla_e(\mu) \geq pla_e(\mu')$ for all $\mu' \in M_e$, gives us that $\mu \in Max_e(M_e) \neq \emptyset$.

[20] These facts hold semantically, for propositions represented as *sets* $P \subseteq M_O$. As we'll see later, there is a difference between update and conditioning at the *syntactic* level. For qualitative AGM updates, this fact is well-known in Dynamic Epistemic Logic.

The rest of the proof goes exactly as in the proof of Proposition 4, by replacing unconditional belief *B(P)*, plausibility *pla* and *Max(M)* by their conditional versions *B(P|e)*, pla_e and $Max_e(M_e)$. □

5 Safe belief, statistical knowledge, and verisimilitude

Until now, we only used the notion of knowledge K that is most common among logicians, economists and computer scientists: absolutely certain, infallible, irrevocable, and fully introspective knowledge. This matches what philosophers call "(hard) evidence" or "(hard) information". But the notion of knowledge favoured by epistemologists is *softer*: fallible, less-than-absolutely-certain, revisable, and possibly non-introspective (or at least not always negatively introspective). It is the kind of knowledge that we typically encounter in daily life or in empirical sciences, where absolute certainty may be hard to achieve. This is known sometimes as *defeasible knowledge*, and it is also related to the notion of *inductive knowledge* in Philosophy of Science. Here, we are interested in developing such a soft notion of knowledge that can apply to *statistical learning*: after repeatedly updating our beliefs by sampling from an unknown distribution, when do our beliefs become focused enough and stable enough to qualify as soft 'knowledge' of the true distribution (at least to some good enough approximation)?

Various formalizations have been proposed for this notion. Here, we will borrow ideas from the so-called Defeasibility Theory of Knowledge (Lehrer 1990): the main principle is that 'knowledge' is a form of robust belief, namely belief that is resilient under conditioning with truthful information. These ideas go back to Plato's *Meno* and were more recently championed in various forms by Klein, Lehrer, Pappas and Swayn, Rott and others. Before going on to formalize and then criticize the defeasibility theory, Stalnaker (1996) summarizes it as follows: "An agent knows that ϕ if and only if ϕ is true, she believes that ϕ, and she continues to believe ϕ if any true information is received". Rott (2004) develops a version called *stability theory*, and states it as: "A belief K is a piece of knowledge of the subject S iff K is not given up by S on the basis of any true information that S might receive". Baltag and Smets (2008b) restated Stalnaker's formalization, under the name of *safe belief*, and developed it in the framework of dynamic epistemic logic. Here, we adapt this concept to our setting, and later strengthen it to a notion of *statistical knowledge*.

Definition 8 [*Safe Belief*] Let $\mathbf{M} = (M, pla)$ be a plausibility model, in which we also specify the 'true' distribution μ. We say that a proposition $P \subseteq M$ is *safely believed* (or is a "safe belief") at μ in $\mathbf{M}$, and write $\mu \models_{\mathbf{M}} Sb(P)$, if P is believed in $\mathbf{M}$ conditional on every *true* proposition Q; i.e. *B(P|Q)* holds for all $Q \in Prop$ with $\mu \in Q$.

This is simply the same notion as the one defined by Baltag and Smets (2008b) in general plausibility models, but stated here in the special case of our probabilistic plausibility models. As such, it satisfies the following general characterization, given in Baltag and Smets (2008b):

Proposition 12 *The following are equivalent:*

- *P is safely believed at μ in* $\mathbf{M}$*;*
- *all distributions in M that are at least as plausible as μ satisfy P; i.e., we have that* $\{\nu \in M \mid pla(\nu) \geq pla(\mu)\} \subseteq P$.

It is easy to see that, if P is a safe belief, then P is a true belief. As such, the notion of safe belief gives a good formal approximation of the defeasibility conception of knowledge.

Distance from the truth and verisimilitude We can think of a plausibility model $\mathbf{M} = (M, pla)$ as an epistemic/doxastic approximation of some unknown probability distribution $\mu \in M$. The natural question that arises is: how 'truthlike' is our model $\mathbf{M}$, how good an approximation is it? To assess this, we connect with notions from Verisimilitude Theory, cf Popper (1976), Tichy (1974), Miller (1974), Niiniluoto (1987, 1998), Kuipers (1987) and others. In particular, we adapt to our setting ideas coming from the metric approach to truthlikeness Niiniluoto (1987). We are looking for a notion of *distance of a model* $\mathbf{M}$ *from a distribution* $\mu \in M$, which measures how far the agent's beliefs are from the truth. In the case of closed models, the beliefs are given by the set $Max(M)$, so the natural notion of distance would be given in this case by the quantity

$$\delta_\mu(\mathbf{M}) := sup\{d(\mu, \nu) \mid \nu \in Max(M)\},$$

which measures the "worst possible error" one could make when taking as the true distribution to be any of the ones compatible with the agent's beliefs. However, when M is *not* closed, we might have $Max(M) = \emptyset$, which would render the above notion of distance-from-the-truth meaningless, or at least useless (in case we adopt the natural convention that $sup\emptyset = \infty$). But one can weaken the above definition to include in the relevant set of possibilities (whose distances from the truth are assessed) all the "plausible enough" distributions, and in particular *all the ones that are at least as plausible as the true distribution*. In this way, we arrive at the following definition of distance-from-the-truth:

$$d_\mu(\mathbf{M}) := sup\{d(\mu, \nu) \mid \nu \in M, pla(\nu) \geq pla(\mu)\} = sup\{d(\mu, \nu) \mid \nu \in M, \nu \geq^{\mathbf{M}} \mu\}.$$

This measures the worst possible error one could make when taking as the true distribution any of the ones that are currently thought to be at least plausible as the "truly true" distribution μ. It is easy to see that we have that the distance-from-the-truth matches the radius of the smallest open ball around the true distribution that is safely believed:

$$d_\mu(\mathbf{M}) = inf\{\varepsilon > 0 \mid \mu \models_{\mathbf{M}} Sb(\mathcal{B}_\varepsilon(\mu))\} = min\{\varepsilon \geq 0 \mid \mu \models_{\mathbf{M}} Sb(\mathcal{B}_\varepsilon(\mu))\}.$$

So $d_\mu(\mathbf{M}) \geq \varepsilon$ tell us that the agent has a *safe belief of the approximate value of the true distribution within an ε-margin of error*. It is also easy to see that we have

$$0 \leq \delta_\mu(\mathbf{M}) \leq d_\mu(\mathbf{M}) \quad \text{whenever } Max(M) \neq \emptyset,$$

and also that we have

$$d_\mu(\mathbf{M})=0 \text{ iff } \delta_\mu(\mathbf{M})=0 \text{ iff } Max(M)=\{\mu\} \text{ iff } \mathbf{M} \models B(\{\mu\}) \text{ iff } \mu \models_{\mathbf{M}} Sb(\{\mu\}).$$

So 0-distance (according to either definition) indicates that the agent's (safe) beliefs fully match the true distribution.

When we have $d_\mu(\mathbf{M}) < d_\mu(\mathbf{M}')$ for the true distribution μ, we say that $\mathbf{M}$ is *more truthlike* than $\mathbf{M}'$. This verisimilitude order suffices for our purposes. But we could also convert it into an actual measure of truthlikeness, by defining the verisimilitude $v_\mu(\mathbf{M})$ of a model $\mathbf{M}$ wrt a distribution μ, say by putting $v_\mu(\mathbf{M}) := 2^{-d_\mu(\mathbf{M})}$. The maximum verisimilitude $v_\mu(\mathbf{M}) = 1$ is achieved when $d_\mu(\mathbf{M}) = 0$, i.e. when $\mu \models_{\mathbf{M}} Sb(\{\mu\})$.

Safe belief is not safe from conditioning on events While of inherent interest, the notion of safe belief does not fully capture the intended meaning of defeasible knowledge in a probabilistic framework. Although safe belief is resilient under conditioning with any true 'proposition', in our setting propositions are not the only kind of new information; and indeed, safe beliefs are *not* necessarily stable under conditioning on events. Indeed, even if we restrict to truly observable events (whose true probability $\mu \neq 0$), one can show that *no non-trivial belief is stable under every such event*!

This means we have to moderate our safety requirements when dealing with events. Note that, for inductive learning, absolute safety (under all observable sampling events) is irrelevant: what is important is that our correct beliefs are resilient *throughout the (actual) future sampling history*. This resembles the notion of identification in the limit in Formal Learning Theory (Gold 1967), as well as the concepts of inductive knowledge developed in e.g. Kelly (2014) and Baltag et al. (2019b). In our setting, this gives rise to the concept of *statistical knowledge*:

Definition 9 [*Statistical Knowledge*] Let $\mathbf{M} = (M, pla)$ be a plausibility model, let μ be some distribution (representing the 'true' probability), and let $\omega \in \Omega$ be an infinite observation stream (representing the 'true' future sampling history from the unknown distribution μ). We say that a proposition $P \subseteq M$ is *statistically known* (or is "statistical knowledge") at μ wrt ω in $\mathbf{M}$, and write $\mu, \omega \models_{\mathbf{M}} Sk(P)$, if P is believed in $\mathbf{M}$ conditional on every 'true' proposition Q and every (event corresponding to an) initial segment of the 'true' sampling history ω; i.e. if we have $B(P|Q, [\omega^{\geq n}])$, for all $Q \in Prop$ with $\mu \in Q$, and all $n \geq 0$.

It is obvious that, if P is statistically known, then it is safely believed. But statistical knowledge is much more resilient: it essentially captures a strong form of inductive knowledge. Using Proposition 12, we immediately obtain the following characterization:

Proposition 13 *The following are equivalent:*

- *P is statistically known at μ wrt ω in $\mathbf{M}$;*
- *after every initial segment $[\omega^{\leq n}]$ of the true sampling history ω, every distribution in M that is at least as plausible as μ satisfies P; i.e. we have:*

$$\forall \nu \in M \, \forall n \geq 0 \left(pla_{[\omega^{\leq n}]}(\nu) \geq pla_{[\omega^{\leq n}]}(\mu) \Rightarrow \nu \in P \right).$$

In the next section, we show that this notion is actually realistically achievable, and in fact unavoidable: repeated sampling will almost surely eventually lead to statistical knowledge of the true distribution with any desired accuracy.

6 Tracking the truth

Definition 10 For $\mu \in M$, we define the set Ω_μ of *μ-normal observations* as the set of infinite sequences from O for which (1) the limiting frequencies of each o_i correspond to $\mu(o_i)$ and (2) no outcome with probability 0 is ever observed:

$$\Omega_\mu := \left\{\omega \in \Omega \mid \forall o \in O \ \lim_{n\to\infty} \frac{|\{i \leq n \mid \omega_i = o\}|}{n} = \mu(o)\}\backslash\{\omega \in \Omega \mid \exists i \in \mathbb{N} \ \mu(\omega_i) = 0\right\}$$

Proposition 14 *For every probability function μ, $\mu(\Omega_\mu) = 1$. Hence, if μ is the true probability distribution over O, then almost all observable infinite sequence from O will be μ-normal.*

Proof Let $\Delta = \{\omega \in \Omega \mid \exists i \in \mathbb{N} \ \mu(\omega_i) = 0\}$. Using the law of large numbers it is enough to show that $\mu(\Delta) = 0$. To see this let $\mu(o) = 0$ then

$$\mu(\{\omega \in \Omega \mid \exists i \in \mathbb{N} \ \omega_i = o\}) = \mu\left(\bigcup_{i\in\mathbb{N}}\{\omega \in \Omega \mid \omega_i = o\}\right) \leq \sum_{i\in\mathbb{N}} \mu(o^i) = 0.$$

The result then follows from finiteness of O. □

We are now in the position to look into the learnability of the correct probability distribution via plausibility-revision induced by repeated sampling. We first prove a preliminary result on convergence.

Lemma 2 *Let $\mathbf{M} = (M, pla)$ be a plausibility model, and $\mu \in M$. Then, when repeatedly sampling from an unknown distribution μ, we have that for every $\varepsilon > 0$, the plausibility of having a distribution ε-farther from μ will become in the limit vanishingly smaller than the plausibility $pla(\mu)$ of the true distribution μ.*

More precisely: for every μ-normal sequence $\omega \in \Omega_\mu$ and every positive real $\varepsilon > 0$, we have

$$\lim_{n\to\infty} \frac{pla_{[\omega^{\leq n}]}(\overline{M}\backslash\mathcal{B}_\varepsilon(\mu))}{pla_{[\omega^{\leq n}]}(\mu)} = 0$$

(where recall that $\mathcal{B}_\varepsilon(\mu) = \{\nu \in M_O | d(\mu, \nu) < \varepsilon\}$).

Proof We first need to make some preliminary notations and observations. If $O = \{o_1, \ldots, o_n\}$ is the set of outcomes, and μ is the fixed distribution in the statement of our Lemma, then we put $p_i := \mu(o_i)$, for all $1 \leq i \leq n$. More generally, for all distributions $\nu \in \overline{M}$, all μ-normal sequences $\omega \in \Omega_\mu$ and all $1 \leq i \leq n$, we put: $\nu_i := \nu(o_i)$, $m_{i,\omega} := |\{k \leq m \mid \omega_k = o_i\}|$ for the number of occurrences of o_i in the sequence $\omega^{\leq m} = (\omega_1, \ldots, \omega_n)$, and $\alpha_{i,m,\omega} := \frac{m_{i,\omega}}{m}$ for the relative frequency of o_i in

$\omega^{\leq m}$. Since $\omega \in \Omega_\mu$) we have (by the definition of Ω_μ) that: $\lim_{m\to\infty} \alpha_{i,m,\omega} = p_i$ for all $1 \leq i \leq n$; and also that $m_{i,\omega} = \alpha_{i,m,\omega} = 0$ holds whenever $p_i = \mu(o_i) = 0$ (because of the normality of the sequence ω).

Let us put $A := \{1 \leq i \leq n \,|\, p_i \neq 0\}$. Since $\lim_{m\to\infty} \alpha_{i,m,\omega} = p_i > 0$ for $i \in A$, there must exist some $N_{1,\omega}$ such that $0 < \frac{p_i}{2} \leq \alpha_{i,m,\omega} \leq 2 \cdot p_i$ for all $m \geq N_{1,\omega}$ and all $i \in A$. Since $0 \leq p_i, \nu_i \leq 1$, this gives us that

$$(*) \quad \nu_i^{m\cdot 2\cdot p_i} \leq \nu_i^{m\cdot\alpha_{i,m,\omega}} \leq \nu_i^{m\cdot\frac{p_i}{2}} \text{ for all } \nu \in \overline{M}, \text{ all } i \in A \text{ and all } m \geq N_{1,\omega},$$

and in particular $p_i^{m\cdot 2\cdot p_i} \leq p_i^{m\cdot\alpha_{i,m,\omega}} \leq p_i^{m\cdot\frac{p_i}{2}}$ for all such ν, i, m.

Using independence, we have

$$(**) \quad pla_{[\omega^{\leq m}]}(\nu) = pla(\nu)\cdot\nu([\omega^{\leq m}]) = pla(\nu)\cdot\Pi_{i=1}^{n}\nu_i^{m_{i,\omega}} = pla(\nu)\cdot\Pi_{i\in A}\nu_i^{m\cdot\alpha_{i,m,\omega}}$$

(where we used the fact that, for every $i \notin A$ we have $p_i = 0$, so by normality of the sequence we also have $m_{i,\omega} = \alpha_{i,m,\omega} = 0$, and thus $\nu_i^{m_{i,\omega}} = 1$, hence these factors can be skipped from the product). In particular, for $\nu := \mu$ (so $\nu_i = p_i$), we get that

$$(***) \quad pla_{[\omega^{\leq m}]}(\mu) = pla(\mu)\cdot\mu([\omega^{\leq m}]) = pla(\mu)\cdot\Pi_{i\in A}p_i^{m\cdot\alpha_{i,m,\omega}} > 0$$

(since $p_i \neq 0$ for $i \in A$, and also $pla(\mu) \neq 0$ because $\mu \in M$).

Using these abbreviations and facts, we can now prove our lemma. Fix $\omega \in \Omega_\mu$ and $\varepsilon > 0$. To prove the desired conclusion, let now $\nu \in \overline{M}\backslash\mathcal{B}_\varepsilon(\mu)$, and let N be any arbitrarily chosen natural number. Using the above unfoldings (**) and (***) of the definitions of $pla(\overline{M}\backslash\mathcal{B}_\varepsilon(\mu))$ and $pla_{[\omega^{\leq n}]}(\mu)$, we see that it is enough to show that, for any such arbitrarily chosen N, we have

$$N\cdot pla(\nu)\cdot\Pi_{i\in A}\nu_i^{m\cdot\alpha_{i,m,\omega}} < pla(\mu)\cdot\Pi_{i\in A}p_i^{m\cdot\alpha_{i,m,\omega}} \tag{1}$$

for all large enough m.

We prove this by cases. In the first case, assume that $pla(\nu) = 0$, then the left hand side of (1) is 0 and the inequality holds. In the second case, assume that $pla(\nu) > 0$. Let $\Delta = \{\nu \in \overline{M} \,|\, \nu_i = 0 \text{ for some } i \in A\}$, and similarly for any $\delta > 0$, put $\Delta_\delta = \{\nu \in \overline{M} \,|\, \nu_I < \delta \text{ for some } i \in A\}$, and so $\overline{\Delta_\delta} = \{\nu \in \overline{M} \,|\, \nu_I \leq \delta \text{ for some } i \in A\}$ is its closure. Choose some $\delta > 0$ small enough such that we have $\Pi_{i\in A}\nu_i^{\frac{p_i}{2}} < \Pi_{i\in A}p_i^{2\cdot p_i}$ for all $\nu \in \overline{\Delta_\delta}$ (-this is possible, since $\Pi_{i\in A}\nu_i^{\frac{p_i}{2}} = 0 < \Pi_{i\in A}p_i^{2\cdot p_i}$ for all $\nu \in \Delta$, so the continuity of $\Pi_{i\in A}\nu_i^{\frac{p_i}{2}}$ gives us the existence of δ). Hence, we have

$$0 \leq \frac{\Pi_{i\in A} v_i^{\frac{p_i}{2}}}{\Pi_{i\in A} p_i^{2\cdot p_i}} < 1 \quad \text{for all } \nu \in \overline{\Delta_\delta}$$

(where we used again the fact that $p_i > 0$ for $i \in A$). The set $\overline{\Delta_\delta}$ is closed, hence the continuous function $\frac{\Pi_{i\in A} v_i^{\frac{p_i}{2}}}{\Pi_{i\in A} p_i^{2\cdot p_i}}$ has a maximum value Q on $\overline{\Delta_\delta}$. Note that $Q < 1$ (-this follows from the inequality above), so there exists some $N_2 > N_{1,\omega}$ (where $N_{1,\omega}$ is the number satisfying the inequality (*) in the preliminary facts above) s.t. we have $Q^m < \frac{pla(\mu)}{N}$ for all $m > N_2$. Recalling also that by definition $pla(\nu) \leq 1$, we obtain, for all $\nu \in \Delta_\delta$:

$$\begin{aligned} &N \cdot pla(\nu) \cdot \Pi_{i\in A} v_i^{m\cdot\alpha_{i,m,\omega}} \leq N \cdot 1 \cdot \Pi_{i\in A} v_i^{m\cdot\frac{p_i}{2}} \leq N \cdot (Q \cdot \Pi_{i\in A} p_i^{2\cdot p_i})^m \\ &= N \cdot Q^m \cdot \Pi_{i\in A} p_i^{m\cdot 2\cdot p_i} < N \cdot \frac{pla(\mu)}{N} \cdot \Pi_{i\in A} p_i^{m\cdot\alpha_{i,m,\omega}} = pla(\mu) \cdot \Pi_{i\in A} p_i^{m\cdot\alpha_{i,m,\omega}} \end{aligned}$$

(where we used the above facts as well as the inequality (*)). So we proved that the inequality (1) holds for all $\nu \in \Delta_\delta$. It thus remains only to prove it for all $\nu \in M' := \overline{M}\backslash(B_\varepsilon(\mu) \cup \Delta_\delta)$. For this, note that $M' := \overline{M}\backslash(B_\varepsilon(\mu) \cup \Delta_\delta)$ is closed and that $\nu_i \neq 0$ over this set for all $i \in A$, while for all $i \notin A$ we have $\alpha_{i,m,\omega} = 0$. Hence using the assumption that $pla(\nu) \neq 0$, (1) is equivalent over this set with:

$$\left(\frac{pla(\mu)}{pla(\nu)}\right) \cdot \left(\frac{\Pi_{i=1}^n p_i^{m\cdot\alpha_{i,m,\omega}}}{\Pi_{i=1}^n \nu_i^{m\cdot\alpha_{i,m,\omega}}}\right) > N \tag{2}$$

Applying logarithm (and using its monotonicity, and its other properties), this in turn is equivalent to

$$\log(pla(\mu)) - \log(pla(\nu)) + \sum_{i=1}^{n} m \cdot \alpha_{i,m,\omega} \cdot (\log p_i - \log \nu_i) > logN \tag{3}$$

So we see that it is enough to show that, for all large m and for $\nu \in M'$, we have

$$m > \frac{logN + log(pla(\nu)) - log(pla(\mu))}{\sum_{i=1}^n \alpha_{i,m,\omega} \cdot (\log p_i - \log \nu_i)} \tag{4}$$

Recall that $\alpha_{i,m,\omega} \geq \frac{p_i}{2}$ for all $m > N_2 > N_1$ and all $1 \leq i \leq n$. Thus, to prove (4), it is enough to show that, for large m and for all $\nu \in M'$, we have

$$m > \frac{f(\nu)}{g(\nu)}, \tag{5}$$

where we introduced the auxiliary continuous functions $f, g : M' \to R$, defined by putting $f(\nu) = 2 \cdot (logN + \log(pla(\nu)) - \log(pla(\mu)))$ and $g(\nu) = \sum_{i=1}^n p_i \cdot (\log p_i - \log \nu_i)$ for all $\nu \in M_O$.

To show (5), note first that

$$g(\nu) = \sum_{i=1}^{n} p_i \cdot (\log p_i - \log \nu_i) = log\left(\frac{\Pi_{i=1}^{n} p_i^{p_i}}{\Pi_{i=1}^{n} \nu_i^{p_i}}\right) > log1 = 0$$

(where at the end we used the fact, proved in Lemma 1, that the measure μ, with values $\mu(o_i) = p_i$, is the unique maximizer of the function $\Pi_{i=1}^{n} \nu_i^{p_i}$ on M_O). Since g is continuous and M' is closed, g is bounded and attains its infimum $B = min_{M'}(g)$ on M'. But since g is non-zero on M', this minimum cannot be zero: $B = min_{M'}(g) \neq 0$. Similarly, since f is continuous and M' is closed, g is bounded and attains its supremum $C = max_{M'}(f) < \infty$ (which thus has to be finite). Take now some $N_3 \geq max(N_2, \frac{C}{B})$. For all $m > N_3$, we have

$$m > \frac{C}{B} \geq \frac{f(\nu)}{g(\nu)}$$

for all $\nu \in M'$, as desired. □

We can now establish our first convergence result.

Theorem 2 [Convergence in plausibility] *Let* $\mathbf{M} = (M, pla)$ *be a plausibility model. If* $\mu \in M$ *is the 'true' distribution, then we have the following:*

1. *when repeatedly sampling from the unknown distribution* μ*, we have that for every* $\varepsilon > 0$*, the plausibility* $pla(\overline{M} \backslash \mathcal{B}_\varepsilon(\mu))$ *of having a distribution* ε*-farther from* μ *will also almost surely converge to* 0 *(as sample size converges to infinity):*

$$\mu(\{\omega \in \Omega \mid \lim_{n\to\infty} pla_{[\omega^{\leq n}]}(\overline{M} \backslash \mathcal{B}_\varepsilon(\mu)) = 0\}) = 1;$$

in particular, in the same conditions of repeated sampling, every other *distribution* $\nu \in \overline{M} \backslash \{\mu\}$ *will almost surely converge to* 0 *(as sample size converges to infinity):*

$$\mu(\{\omega \in \Omega \mid \lim_{n\to\infty} pla_{[\omega^{\leq n}]}(\nu) = 0\}) = 1;$$

2. *in contrast, in the same conditions, we have that for every* $\varepsilon > 0$*, the plausibility* $pla(\mathcal{B}_\varepsilon(\mu))$ *of having a distribution* ε*-close to* μ *will also almost surely eventually settle on* 1 *(after finitely many rounds of sampling):*

$$\mu(\{\omega \in \Omega \mid \exists N \forall n \geq N pla_{[\omega^{\leq n}]}(\mathcal{B}_\varepsilon(\mu)) = 1\}) = 1;$$

as an obvious consequence, in the same conditions, we have for every $\varepsilon > 0$*, the plausibility* $pla(\mathcal{B}_\varepsilon(\mu))$ *of having a distribution* ε*-close to* μ *will also almost surely converge to* 1*:*

$$\mu\left(\left\{\omega \in \Omega \mid \lim_{n\to\infty} pla_{[\omega^{\leq n}]}(\mathcal{B}_\varepsilon(\mu)) = 1\right\}\right) = 1;$$

Proof Fix $\mu \in M$. It is obviously enough to show the following two claims, for all $\mu - normal$ sequences $\omega \in \Omega_\mu$ and all $\varepsilon > 0$:

$$\lim_{n\to\infty} pla_{[\omega^{\leq n}]}(\overline{M}\backslash\mathcal{B}_\varepsilon(\mu)) = 0 \text{ for all } \varepsilon > 0;$$
$$\exists N \forall n \geq N \; pla_{[\omega^{\leq n}]}(\mathcal{B}_\varepsilon(\mu)) = 1.$$

To prove the first claim, we use the fact that every plausibility ranking function satisfies $0 \leq pla \leq 1$ to derive

$$0 \leq pla_{[\omega^{\leq n}]}(\overline{M}\backslash\mathcal{B}_\varepsilon(\mu)) \leq \frac{pla_{[\omega^{\leq n}]}(\overline{M}\backslash\mathcal{B}_\varepsilon(\mu))}{pla_{[\omega^{\leq n}]}(\mu)} \cdot pla_{[\omega^{\leq n}]}(\mu)$$
$$\leq \frac{pla_{[\omega^{\leq n}]}(\overline{M}\backslash\mathcal{B}_\varepsilon(\mu))}{pla_{[\omega^{\leq n}]}(\mu)} \cdot 1 = \frac{pla_{[\omega^{\leq n}]}(\overline{M}\backslash\mathcal{B}_\varepsilon(\mu))}{pla_{[\omega^{\leq n}]}(\mu)},$$

then obtain the desired conclusion by taking limits and applying Lemma 2.

For the second claim: for any $\epsilon > 0$, apply the first claim to conclude that $\exists N \forall n \geq N \; pla_{[\omega^{\leq n}]}(\overline{M}\backslash\mathcal{B}_\varepsilon(\mu)) \leq \frac{1}{2}$. From this we get that

$$1 = pla_{[\omega^{\leq n}]}(M) = max(pla_{[\omega^{\leq n}]}(M \cap \mathcal{B}_\varepsilon(\mu)\}, pla_{[\omega^{\leq n}]}(M\backslash\mathcal{B}_\varepsilon(\mu)\})$$
$$\leq max(pla_{[\omega^{\leq n}]}(M \cap \mathcal{B}_\varepsilon(\mu)\}, \frac{1}{2}),$$

hence $pla_{[\omega^{\leq n}]}(M \cap \mathcal{B}_\varepsilon(\mu)\} = 1$, and so also $pla_{[\omega^{\leq n}]}\mathcal{B}_\varepsilon(\mu) = 1$. □

Corollary 1 [Convergence in belief] *Let* $\mathbf{M} = (M, pla)$ *be a plausibility model. Then the agent's beliefs after repeated sampling will almost surely eventually settle arbitrarily close to the true distribution.*

More precisely: for every $\mu \in M$ *and every* $\varepsilon > 0$*, we have*

$$\mu(\{\omega \in \Omega \mid \exists N \; \forall n \geq N \quad B(\mathcal{B}_\varepsilon(\mu)) \textit{ holds in } \mathbf{M}_{[\omega^{\leq n}]}\}) = 1,$$

or equivalently

$$\mu(\{\omega \in \Omega \mid \exists N \; \forall n \geq N \quad B(\mathcal{B}_\varepsilon(\mu) \mid [\omega^{\leq n}]) \textit{ holds in } \mathbf{M}\}) = 1.$$

Proof From Theorem 2, we know that with μ-probability 1, we have $\lim_{n\to\infty} pla_{[\omega^{\leq n}]}(\overline{M}\backslash\mathcal{B}_\varepsilon(\mu)) = 0$, hence almost certainly there is some N_1 such that $pla_{[\omega^{\leq n}]}(\overline{M}\backslash\mathcal{B}_\varepsilon(\mu)) < 1$ for all $n \geq N_1$. Similarly, we know from Theorem 2 that, with μ-probability 1, there is some N_2 such that $pla_{[\omega^{\leq n}]}(\mathcal{B}_\varepsilon(\mu)) = 1$ for all $n \geq N_2$. By taking $N := max\{N_1, N-2\}$, we obtain that (with μ-probability 1): for all $n \geq N$ and all $\nu \in M$ with maximal plausibility $pla_{[\omega^{\leq n}]}(\nu) = pla_{[\omega^{\leq n}]}(\nu)(M) = 1$, we have $\nu \in \mathcal{B}_\varepsilon(\mu)$, as desired. □

We now show that we can strengthen this result to:

Proposition 15 [Convergence in safe belief] *Let* $\mathbf{M} = (M, pla)$ *be a plausibility model. Then the agent's safe beliefs after repeated sampling will almost surely eventually settle arbitrarily close to the true distribution.*

More precisely: for every $\mu \in M$ *and every* $\varepsilon > 0$*, we have*

$$\mu(\{\omega \in \Omega \mid \exists N\ \forall n \geq N \quad Sb(\mathcal{B}_\varepsilon(\mu)) \text{ holds at } \mu \text{ in } \mathbf{M}_{[\omega^{\leq n}]}\}) = 1.$$

Proof Let $\omega \in \Omega_\mu$. By Lemma 2, we have $\lim_{n\to\infty} \frac{pla_{[\omega^{\leq n}]}(\overline{M}\backslash\mathcal{B}_\varepsilon(\mu))}{pla_{[\omega^{\leq n}]}(\mu)} = 0$. So there exists some N, s.t. $\frac{pla_{[\omega^{\leq n}]}(\overline{M}\backslash\mathcal{B}_\varepsilon(\mu))}{pla_{[\omega^{\leq n}]}(\mu)} < 1$ for all $n \geq N$. Hence, we have $pla_{[\omega^{\leq n}]}(\overline{M}\backslash\mathcal{B}_\varepsilon(\mu)) < pla_{[\omega^{\leq n}]}(\mu)$ for all $n \geq N$. Thus, for all $n \geq N$ and all $\nu \in M$, if $pla_{[\omega^{\leq n}]}(nu) \geq_{[\omega^{\leq n}]} (\mu)$, then $\nu \notin \overline{M}\backslash\mathcal{B}_\varepsilon(\mu))$, i.e. $\nu \in \mathcal{B}_\varepsilon(\mu))$. By Proposition 12, this means that $Sb(P)$ holds in μ in $\mathbf{M}_{[\omega^{\leq n}]}$ for all $n \geq 1$. The desired conclusion follows again from the fact that $\mu(\Omega_\mu) = 1$. □

An obvious consequence is the following:

Corollary 2 [Approximate statistical learning] *Let* $\mathbf{M} = (M, pla)$ *be a plausibility model. Then after repeated sampling from an unknown distribution, the agent will almost surely eventually acquire approximate statistical knowledge of the true distribution with any desired accuracy* $\varepsilon > 0$*.*

More precisely: for every $\mu \in M$ *and every* $\varepsilon > 0$*, we have*

$$\mu(\{\omega \in \Omega \mid \exists N \quad Sk(\mathcal{B}_\varepsilon(\mu)) \text{ holds at } \mu \text{ wrt } \omega^{\geq N} \text{ in } \mathbf{M}_{[\omega^{\leq N}]}\}) = 1.$$

The proof is immediate, given Proposition 15. All these convergence results are inexact: they concern only *approximations* of the true distribution. However, the fact that every non-zero degree of accuracy is eventually achieved (and maintained forever after) shows that the verisimilitude of our models keeps increasing, or equivalently the distance-from-the-truth keeps decreasing (approaching 0 in the limit). In this sense, we have convergence *in the limit* to the *exact* true distribution:

Corollary 3 [Convergence in verisimilitude] *Let* $\mathbf{M} = (M, pla)$ *be a plausibility model. If* $\mu \in M$ *is the true distribution, then the distance-from-the-truth will almost surely converge to* 0 *after repeated sampling:*

$$\mu(\{\omega \in \Omega \mid lim_{n\to\infty} d_\mu(\mathbf{M}_{[\omega^{\leq n}]}) = 0\}) = 1.$$

Proof It is clear that we have to show that

$$\mu(\{\omega \in \Omega \mid \forall \varepsilon > 0 \exists N \forall n \geq N\ d_\mu(\mathbf{M}_{[\omega^{\leq n}]}) < \epsilon\}) = 1.$$

But note that, by the definition of distance-to-the-truth, we have the following equivalence:

$$d_\mu(\mathbf{M}_{[\omega^{\leq n}]}) < \epsilon \text{ iff } Sb(\mathcal{B}_\varepsilon(\mu)) \text{ holds at } \mu \text{ in } \mathbf{M}_{[\omega^{\leq n}]}.$$

The desired conclusion follows immediately, given Proposition 15. □

A general feature of all the above forms of truth-tracking is that the convergence to the *exact* true distribution (rather than to an approximation) happens *only in the limit* (rather than being reached at some finite stage). However, one can do better than this when the agent's prior knowledge is consistent with only a *discrete* (or in particular, a *finite*) set of distributions:

Proposition 16 [Finite convergence to exact truth] *Let* $\mathbf{M} = (M, pla)$ *be a plausibility model, based on a discrete set* $M \subseteq M_O$.[21] *Then we have the following:*

– *when repeatedly sampling from an unknown distribution* μ*, the plausibility* $pla(\mu)$ *of the true distribution will almost surely eventually settle on* 1 *(after finitely many rounds of sampling); while in contrast, the plausibility* $pla(\nu)$ *of any* other *distribution will almost surely settle* below *any given threshold* $\delta > 0$ *(after finitely may such rounds):*

$$\mu(\{\omega \in \Omega \mid \exists N \forall n \geq N \; pla_{[\omega^{\leq n}]}(\mu) = 1\}) = 1, \quad \textit{and}$$
$$\mu(\{\omega \in \Omega \mid \exists N \forall n \geq N \; pla_{[\omega^{\leq n}]}(\nu) < \delta\}) = 1, \; \textit{for all } \nu \neq \mu \textit{ and all } \delta > 0;$$

– *similarly, the agent's beliefs will almost surely eventually settle on the exact true probability* μ*, after finitely many rounds sampling:*

$$\mu(\{\omega \in \Omega \mid \exists N \forall n \geq N \; B(\{\mu\}) \textit{ holds in } \mathbf{M}_{[\omega^{\leq n}]}\}) = 1;$$

– *the same statement as in the previous part applies to safe beliefs:*

$$\mu(\{\omega \in \Omega \mid \exists N \forall n \geq N \; Sb(\{\mu\}) \textit{ holds at } \mu \textit{ in } \mathbf{M}_{[\omega^{\leq n}]}\}) = 1;$$

– *after finitely many rounds of sampling from the unknown distribution, the agent will almost surely eventually acquire exact statistical knowledge of the true distribution:*

$$\mu(\{\omega \in \Omega \mid \exists N \forall n \geq N Sk(\{\mu\}) \text{ holds at } \mu \text{ wrt } \omega^{\geq N} \text{ in } \mathbf{M}_{[\omega^{\leq n}]}\}) = 1.$$

– *finally, the distance-to-the-truth of the plausibility model will almost surely eventually settle to* 0*, after finitely many rounds of sampling:*

$$\mu(\{\omega \in \Omega \mid \exists N \forall n \geq N \; d_\mu(\mathbf{M}_{[\omega^{\geq n}]}) = 0\}) = 1.$$

Proof Apply each of the previous results to some $\varepsilon > 0$ small enough so that $\mathcal{B}_\varepsilon(\mu) \cap M = \{\mu\}$. □

[21] As usual in topology, a set $M \subseteq M_O$ of distributions is *discrete* (wrt the standard topology) if every distribution $\mu \in M$ in the set is an *isolated point*, i.e. it has a neighborhood $\mathcal{B}_\varepsilon(\mu)$, with $\varepsilon > 0$ and $M \cap \mathcal{B}_\varepsilon(\mu) = \{\mu\}$. Every finite set $M \subseteq M_O$ is discrete. An example of an infinite discrete set is obtained by taking in the Coin Example the set M of all distributions assigning to Heads a probability of the form $\frac{1}{n}$, for any natural number $n > 0$.

It is important to note the differences between our convergence results and the Savage-style convergence results in the Bayesian literature (Edwards et al. 1963; Savage 1954; Doob 1971; Gaifman and Snir 1982; Earman 1992), that were mentioned in the Introduction. *Savage's theorem assumes a certain restriction on the true hypothesis* (namely, that its prior probability is non-zero), *which makes it applicable only to a finite (or countable) set of hypotheses*[22] (since otherwise the prior probability cannot be assumed to be non-zero for every hypothesis). Our general results (concerning truth-tracking in the limit) do not need this assumption and indeed, they even apply to the whole (uncountable) set M_O of all distributions.

On the other hand, in the case of a *finite* (or more generally, *discrete*) set of hypotheses/distributions, our plausibilistic learning is even better-behaved than the standard Bayesian learning: we obtain convergence in this case in *finitely* many steps (while Savage's still converges only in the limit). This faster convergence is explained by the qualitative nature of our belief-formation (as standard in logic, *only the most plausible hypotheses* matter for beliefs), instead of the quantitative-cumulative of probabilistic credences. The combination of this qualitative-logical way of forming beliefs with the statistical-Bayesian way of updating them (as encoded in our rule for conditioning on events) ensures that the true distribution will eventually reach the highest plausibility (among a finite set of distributions), thus giving us finite convergence to the exact truth.

7 Towards a logic of statistical learning

In this section we propose a logical setting that can capture the dynamics of statistical learning described in this paper. Our logical language is designed to accommodate both types of information, i.e. finite observations and higher-order information. As already mentioned, there is a fundamental distinction between these two types of information. The observations are interpreted in a σ-algebra $\mathcal{E} \subseteq \mathcal{P}(\Omega)$, and are not themselves formulas in our formal logical language, as they do not correspond to properties of probability distributions. The formulas will instead be statements *about* the probabilities of observations, given in terms of linear inequalities and logical combinations thereof, as well as the statements concerning the dynamics arising from finite observations.

Given the set of outcomes $O = \{o_1, \ldots, o_n\}$, the set of formulas ϕ of our language is inductively defined by

$$\phi ::= \sum_{i=1}^{m} a_i P(\omega_i) \geq c \mid \phi \wedge \phi \mid \neg\phi \mid K\phi \mid Sb(\phi) \mid B(\phi \mid \omega^{\leq n}) \mid [o]\phi \mid [\phi]\phi$$

where $o, \omega_i \in O$, a_i's and c in $\mathbb{Q}$ and $\omega^{\leq n} = (\omega_1, \ldots, \omega_n) \in O^n$ is a stream of observations of length n.

Let $\mathbf{M} = (M, pla)$ be a probabilistic plausibility model. The semantics is given by inductively defining a satisfaction relation $\mathbf{M}, \mu \vDash \phi$ between distributions $\mu \in M$

[22] This would correspond in our setting to a finite or countable set M.

and formulas ϕ. At each pair $(\mathbf{M}, \mu)$, the symbol P will be interpreted as a probability mass function, namely μ itself. In this definition, we use the notation $\|\phi\|_{\mathbf{M}} := \{\mu \in M \mid \mathbf{M}, \mu \vDash \phi\}$, and skip the subscript $\mathbf{M}$ when the model is understood:

$$
\begin{array}{ll}
\mathbf{M}, \mu \vDash \sum_{i=1}^{n} a_i P(\omega_i) \geq c & \Longleftrightarrow \sum_{i=1}^{n} a_i \mu(\omega_i) \geq c \\
\mathbf{M}, \mu \vDash \phi \wedge \psi & \Longleftrightarrow \mathbf{M}, \mu \vDash \phi \text{ and } \mathbf{M}, \mu \vDash \psi \\
\mathbf{M}, \mu \vDash \neg\phi & \Longleftrightarrow \mathbf{M}, \mu \nvDash \phi \\
\mathbf{M}, \mu \vDash K\phi & \Longleftrightarrow \mathbf{M}, \nu \vDash \phi \text{ for all } \nu \in M \\
\mathbf{M}, \mu \vDash Sb\phi & \Longleftrightarrow \mathbf{M}, \mu \vDash \phi \text{ for all } \nu \in M \text{ s.t. } pla(\nu) \geq pla(\mu) \\
\mathbf{M}, \mu \vDash B(\phi \mid \omega^{\leq n}) & \Longleftrightarrow B(\|\phi\| \mid [\omega^{\leq n}]) \text{ holds in } \mathbf{M} \\
\mathbf{M}, \mu \vDash [o]\phi & \Longleftrightarrow \big(\mu(o) > 0 \implies \mathbf{M}_{[o^1]}, \mu \vDash \phi\big) \\
\mathbf{M}, \mu \vDash [\theta]\phi & \Longleftrightarrow \big(\mathbf{M}, \mu \vDash \theta \implies \mathbf{M}_{\|\theta\|}, \mu \vDash \phi\big)
\end{array}
$$

The atomic formulas $\sum_{i=1}^{m} a_i P(\omega_i) \geq c$ describe linear inequalities satisfied by the true probability, using numerical constants ranging over rationals. The propositional connectives $\neg$, $\wedge$ are standard. Letters K and B stand for knowledge and (conditional) belief operators, and Sb stands for safe belief. The dynamic modalities $[o]\psi$ (standing for "after observing o, ψ holds") and $[\phi]\psi$ (standing for "after learning ϕ, ψ holds") capture the *updates* induced by the two forms of learning.

The reason we did *not* include simple belief $B\phi$ or propositionally-conditional beliefs $B(\phi \mid \phi)$ is that these operators are *definable* as abbreviations in the above syntax. For plain belief, it should be obvious that it can be obtained as a special case of conditioning on a sampling sequence $\omega^{\leq 0}$ of length 0, i.e. we can put

$$B(\phi) \;:=\; B(\phi \mid \lambda),$$

where $\lambda = () = \omega^0$ is the empty sequence of observations. Less trivially, conditional beliefs of the form $B(\phi \mid \theta)$ can be defined in terms of knowledge and safe belief, by putting:

$$B(\phi \mid \theta) \;:=\; \tilde{K}\theta \to \tilde{K}(\theta \wedge Sb(\theta \to \phi)),$$

where $\tilde{K}\psi := \neg K \neg \psi$ is the Diamond-dual modality for K (denoting "epistemic possibility"). With these abbreviations, one can easily check that the resulting notion satisfies the expected semantic clause for conditional belief:[23]

$$\mathbf{M}, \mu \vDash B(\phi \mid \theta) \quad \text{iff} \quad B(\|\phi\| \mid \|\theta\|) \;\text{ holds in }\; \mathbf{M}.$$

We say that a formula ϕ is *valid in model* $\mathbf{M}$, and write $\mathbf{M} \vDash \phi$, if and only if $\mathbf{M}, \mu \vDash \phi$ for all $\mu \in M$. As usual, ϕ is simply *valid* if it is valid in every model $\mathbf{M}$.

Proposition 17 *Let $o \in O$ and formulas ϕ, ψ, θ, ξ. Then the following formulas are valid:*

[23] We mention this fact here without proof, since it is just a special case of a more general observation made in Baltag and Smets (2008b): the above abbreviation matches the semantics of conditional beliefs in any (qualitative) plausibility model based on total preorders $\leq$.

1. $P(o) \geq 0$
2. $\sum_{o \in O} P(o) = 1$
3. $K(\phi \to \theta) \to (K\phi \to K\theta)$
4. $K\phi \to \phi$
5. $K\phi \to KK\phi$
6. $\neg K\phi \to K\neg K\phi$
7. $K\phi \to Sb\phi$
8. $Sb\phi \to \phi$
9. $Sb\phi \to SbSb\phi$
10. $(K(\phi \vee Sb\psi) \wedge K(\psi \vee Sb\phi)) \to (K\phi \vee K\psi)$
11. $B(\phi \to \theta \mid \psi) \to (B(\phi \mid \psi) \to B(\theta \mid \psi))$
12. $K\phi \to B(\phi \mid \psi)$
13. $B(\phi \mid \phi)$
14. $B(\phi \mid \psi) \to K(B(\phi \mid \psi) \mid \psi)$
15. $\neg B(\phi \mid \psi) \to K(\neg B(\phi \mid \psi) \mid \psi)$
16. $B(\theta \mid \phi) \to (B(\xi \mid \phi \wedge \theta) \leftrightarrow B(\xi \mid \phi))$
17. $\neg B(\neg\theta \mid \phi) \to (B(\xi \mid \phi \wedge \theta) \leftrightarrow B(\theta \to \xi \mid \phi))$
18. *If* $\phi \leftrightarrow \theta$ *is valid in* $\mathbf{M}$ *then so is* $B(\xi \mid \phi) \leftrightarrow B(\xi \mid \theta)$.

Proof Note that the plausibility function induces a complete preorder on the set of worlds. The validity of the above formulas over such models follows directly from the results in Board (2004) and Baltag and Smets (2008b), and it is in fact a straightforward application of general results in Correspondence Theory for modal frames. □

Finally, we give some validities regarding the interaction of the dynamic modalities with knowledge modality and (conditional) belief.

Proposition 18 *Let* $o, \omega_1, \ldots, \omega_n \in O$ *and formulas* ϕ, θ, ξ. *Then the following formulas are valid:*

1. $[\phi]q \leftrightarrow (\phi \to q)$ *for atomic* q
2. $[o]q \leftrightarrow (P(o) > 0 \to q)$ *for atomic* q
3. $[\phi]\neg\theta \leftrightarrow (\phi \to \neg[\phi]\theta)$
4. $[o]\neg\theta \leftrightarrow (P(o) > 0 \to \neg[o]\theta)$
5. $[\phi](\theta \wedge \xi) \leftrightarrow ([\phi]\theta \wedge [\phi]\xi)$
6. $[o](\theta \wedge \xi) \leftrightarrow ([o]\theta \wedge [o]\xi)$
7. $[\phi]K\theta \leftrightarrow (\phi \to K[\phi]\theta)$
8. $[o]K\phi \leftrightarrow (P(o) > 0 \to K[o]\phi)$
9. $[\phi]B(\theta \mid \xi) \leftrightarrow (\phi \to B([\phi]\theta \mid \phi \wedge [\phi]\xi))$
10. $[o]B(\phi \mid \omega_1, \ldots, \omega_n) \leftrightarrow (P(o) > 0 \to B([o]\phi \mid o, \omega_1, \ldots, \omega_n))$

Open question. Is the above logic recursively axiomatizable? Is it decidable?
Further extension. To define statistical knowledge, we need to extend the above semantics, by making explicit the actual (future) sampling history. This means that we define the satisfaction relation on triples $\mathbf{M}, \mu, \omega \vDash \phi$, where $\mathbf{M}$ and μ are as above, while $\omega \in \Omega_\mu$ is the infinite string of future observations. The semantical clauses for all the above operators stay essentially the same (i.e. the sequence ω plays no role, so it is just carried through). But we can now introduce new operators, which refer to the

future sampling history. We could directly introduce statistical knowledge $Sk\phi$, but it seems more natural to add instead temporal operators $\Box\phi$ ("from now and forever in the future, ϕ holds") and its dual $\Diamond\varphi$ ("ϕ holds now or at some future moment"), with the obvious semantics:

$$\begin{aligned} \mathbf{M}, \mu, \omega \vDash \Box\phi &\iff \mathbf{M}_{[\omega^{\leq n}]}, \mu, \omega^{>n} \vDash \phi \text{ for all } n \geq 0 \\ \mathbf{M}, \mu, \omega \vDash \Diamond\phi &\iff \mathbf{M}_{[\omega^{\leq n}]}, \mu, \omega^{>n} \vDash \phi \text{ for some } n \geq 0 \end{aligned}$$

(In fact, $\Diamond\phi$ is redundant: it is the Diamond-dual of $\Box$, so can be taken to be just an abbreviation for $\neg\Box\neg\phi$.)

For non-epistemic[24] formulas P, we can identify statistical knowledge $Sk(P)$ with the formula $\Box Sb(P)$. As a result, our result in Corollary 2, on eventual convergence (in finitely many steps) to approximate statistical knowledge of the true distribution, is captured in this logic by the validity

$$(\bigwedge_i P(o_i) = p_i) \rightarrow \Diamond\Box Sb \bigwedge_i (p_i - \epsilon < P(o_i) < p_i + \epsilon),$$

for every $\epsilon > 0$.

8 Conclusion and comparison with other work

We studied forming beliefs about unknown probabilities in situations that are commonly described as those of radical uncertainty. The most widespread approach to model such situations of 'radical uncertainty' is in terms of imprecise probabilities, i.e. representing the agent's knowledge as a set of probability measures. There is extensive literature on the study of imprecise probabilities (Bradley and Drechsler 2014; Chandler 2014; Hajek and Smithson 2012; Levi 1985; Walley 2000; Denoeux 2000; Romeijn and Roy 2014) and on different approaches for decision making based on them Bradley and Steele (2014), Huntley et al. (2014), Troffaesin (2007), Elkin and Wheeler (2018), Mayo-Wilson and Wheeler (2016), Seidenfeld (2004), Seidenfeld et al. (2010), Williams and Robert (2014) or to collapse the state of radical uncertainty by settling on some specific probability assignment as the most rational among all that is consistent with the agent's information. The latter giving rise to the area of investigation known as the Objective Bayesian account (Paris and Rad 2010; Paris and Vencovska 1997; Paris 2014; Rad 2017; Williamson 2008, 2010).

A similar line of inquiry has been extensively pursued in the Economics literature, as well as in Decision Theory, where the situation we are investigating in this paper is referred to as *Knightian uncertainty* or 'ambiguity'. This is the case when the decision-maker has too little information to arrive at a unique prior. There have been different approaches in this literature to model these scenarios. These include, among others, the use of Choquet integration, by for instance Huber and Strassen (1973), or Schmeidler (1989, 1986), the maxmin expected utility by Gilboa and Schmeidler (1989) and the smooth ambiguity model by Klibanoff et al. (2005) which employs second-order

[24] These are formulas that do not contain any of the epistemic operators K, Sb or $B(\phi \mid \omega^{\leq n})$.

probabilities or Al-Najjar's work (Al-Najjar 2009) where he models rational agents who use frequentist models for interpreting the evidence and investigates learning in the long run. Cerreia-Vioglio et al. (2013) studies this problem in a formal setting similar to the one used here and axiomatizes different decision rules such as the maxmin model of Gilboa-Schmeidler and the smooth ambiguity model of Klibanoff et al, and gives an overview of some of the different approaches in that literature.

These approaches employ different mechanisms for ranking probability distributions compared to what we propose in this paper. Among these, it is particularly worth pointing out the difference between our setting and those ranking probability distributions by their (second-order) probabilities. In contrast, in our setting, it is only the worlds with the highest plausibility that play a role in specifying the set of beliefs. In particular, unlike the probabilities, the plausibilities are not cumulative in the sense that the distributions with low plausibility do not add up to form more plausible events as those with low probability would have had. This is a fundamental difference between our account and the account given in terms of second-order probabilities.

Another approach to deal with these scenarios in the Bayesian literature is based on the series of convergence results, that are collectively referred to as "washing out of the prior". The idea, which traces back to Savage, see Edwards et al. (1963) and Savage (1954), is that as long as one repeatedly updates a prior probability for an event through conditionalisation on new evidence, then in the limit one would surely converge to the true probability, independent of the initial choice of the prior.[25] Bayesians use these results to argue that an agent's choice of a probability distribution in scenarios such as our urn example is unimportant as long as she repeatedly updates that choice (via conditionalisation) by acquiring further evidence, for example by repeated sampling from the urn. However, it is clear that the efficiency of the agent's choice for the probability distribution, put in the context of a decision problem, depends strongly on how closely the chosen distribution tracks the actual one. This choice is most relevant when the agents are facing a one-off decision problem, where their approximation of the true probability distribution at a given point ultimately determines their actions at that point.

Our approach, based on forming rational qualitative beliefs *about* probability (based on the agent's assessment of each distribution plausibility), does not seem prone to these objections. The agent does "the best she can" *at each moment*, given her evidence, her higher-order information, and her background assumptions (captured by her plausibility map). Thus, she can solve one-off decision problems to the best of her ability. And, by updating her plausibility with new evidence, her beliefs are still guaranteed to converge to the true distribution (if given enough evidence) in essentially all conditions (including in the cases that evade Savage-type theorems).

[25] To be more precise, if one starts with a prior probability for an event A, and keeps updating this probability by conditionalising on new evidence, then almost surely, the conditional probability of A converges to the indicative function of A (i.e. to 1 if A is true, and to 0 otherwise). This form is called Levy's 0–1 law. Savage's results use IID trials and objective probabilities and have been criticised regarding its applicability to scientific inference. There are, however, a number of more powerful convergence results avoiding these assumptions, for example, based on Doob's martingale convergence theorem (Doob 1971). There are also several generalisations of these results, e.g. Gaifman and Snir (1982).

As we already mentioned, our approach is based on a probabilistic adaptation of the standard qualitative theory of plausibility models (Board 2004; Baltag and Smets 2008b), that underlies modern presentations of standard Belief Revision Theory (Alchourrón et al. 1985; Grove 1988) within Dynamic Epistemic Logic (Baltag and Moss 2004; Baltag et al. 1998; van Ditmarsch et al. 2007; Baltag and Smets 2008b; Baltag and Renne 2016; van Benthem 2011). As such, it has some connections with Wolfgang Spohn's quantitative theory of plausibility ranking (Spohn 2016), but it differs from it in essential ways: like Spohn's ranking theory,[26] we use maximization to form beliefs (where standard probabilistic theory uses addition);[27] but, when updating plausibility with independent sampling evidence, we follow the probabilistic usage of taking products (in Bayes's rule), while Spohn's ranking theory uses addition for this purpose. On the other hand, our framework does satisfy the conditions of Halpern's abstract theory of algebraic conditional plausibility spaces (Halpern 2003), which is meant as a generalization of a large number of theories of uncertainty (Bayesian probabilities, Dempster-Shafer belief functions, possibility measures, relative likelihoods, AGM conditioning, Popper measures, Spohn's ranking theory). The theory postulates the existence of two operations: one, the analogue of probabilistic addition, is used for computing the plausibility of a proposition P, and decide whether it is to be believed or not; while the other, the analogue of probabilistic multiplication, is used for updating plausibilities (via an abstract analogue of Bayes' rule) and for computing the plausibility of joint independent observations. To work well, the two operations need to satisfy certain conditions, tying them together. Our particular combination, of maximization and multiplication, though as far as we know was never encountered in the literature, satisfies Halpern's conditions, and so it is in a sense a "natural" theory. But beyond that, we think that this particular combination is the key to fast learning from sampling, as well as to reconciling probability with logic: on the one hand, multiplication is needed for the update, to deal rationally with successive independent observations (cf. Proposition 9, which would fail without the use of multiplication in our plausibilistic analogue of Bayes' rule); and on the other hand, the use of maximization in the formation of beliefs allows convergence in finitely many steps (in contrast to mere convergence in the limit via probabilistic updating a la Savage), and at the same time makes beliefs about probability fit the general patterns and conditions of Doxastic Logic and Belief Revision Theory. Indeed, it does seem that the particular combination provided by our probabilistic plausibility theory succeeds in adopting the best features of both worlds (doxastic logic with its belief revision, and statistical reasoning with its Bayesian updates), while at the same time fitting within the general conditions of a natural theory of uncertainty (as formalized by Halpern's abstract requirements).

Our approach connects well with mainstream epistemology and formal learning theory, by making essential use of the formal concept of "safe belief", studied in Baltag and Smets (2008b) as an approximation of the philosophical notion of defeasible knowledge (Lehrer 1990; Rott 2004), and related also to the issue of stability or 'resilience' of probabilistic belief (Skyrms 2011), an issue underlying recent attempts

[26] Spohn (2016) uses minimization, but this is only because his setting takes an *implausibility* order as basic.

[27] Witness our normalization condition $sup\{pla(\mu) \mid \mu \in M\} = 1$.

at unifying logical and probabilistic reasoning, cf. the so-called stability theory of belief (Leitgeb 2017). Our concept of statistical knowledge improves on the notion of safe belief, by adding a form of stability under future sampling, that connects well with the learning-theoretic concept of identifiability in the limit (Gold 1967), as well as with various formal notions of inductive knowledge, introduced in Baltag et al. (2019a, b) and Kelly (2014) as epistemic correlatives of empirical induction. As already mentioned, the correlative notion of distance-from-the-truth fits well with the main tenets of Verisimilitude Theory, originating in the work of Popper (1976) and his critics (Tichy 1974; Miller 1974), and developed to maturity in the wor1k of Niiniluoto (1987), Niiniluoto (1998), Kuipers (1987) and others. In particular, our setting fits within the metric approach to truthlikeness (Niiniluoto 1987), resulting in the verisimilitude version of our convergence results: tracking the truth is then naturally understood as progressive increase in our models' truthlikeness (or equivalently, progressive decrease of the models' distance-from-the-truth).

Our paper ends by sketching the contours of a dynamic doxastic logic for statistical learning, that validates a number of standard axioms, and can express the core of our convergence results. Nevertheless, this leads us to an outstanding open problem: finding a complete axiomatization of this logic and investigating its complexity. This seems a daunting task at the time of our writing. Given the power of this formalism and its significance for the investigation of statistical learning, we think this to be an important and potentially fertile challenge.

Acknowledgements We thank the referees for giving us valuable feedback on this paper. During the writing of this paper, Soroush Rafiee Rad's research has been partly funded by the Deutsche Forschungsgemeinschaft (DFG) and the Agence Nationale de la Recherche (ANR) as part of the joint project Collective Attitude Formation [RO 4548/8-1].

References

Alchourrón, C. E., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, *50*, 510–530.

Al-Najjar, N. (2009). Decision makers as statisticians: Diversity, ambiguity, and learning. *Econometrica*, *77*(5), 1339–1369.

Baltag, A., Gierasimczuk, N., Özgün, A., Vargas-Sandoval, A. L., & Smets, S. (2019b). A dynamic logic for learning theory. *Journal of Logical and Algebraic Methods in Programming, 109*.

Baltag, A., Gierasimczuk, N., & Smets, S. (2016). On the solvability of inductive problems: A study in epistemic topology. In *Proceedings of the 15th conference on theoretical aspects of rationality and knowledge* (TARK 2015), ENTCS 215, pp. 81–98.

Baltag, A., Gierasimczuk, N., & Smets, S. (2019a). Truth-tracking by belief revision. *Studia Logica,107,* 917–947.

Baltag, A., & Moss, L. S. (2004). Logics for epistemic programs. *Synthese*, *139*(2), 165–224.

Baltag, A., Moss, L. S., & Solecki, S. (1998). The logic of public announcements, common knowledge, and private suspicions. In I. Gilboa (Ed.), *Proceedings of the 7th conference on theoretical aspects of rationality and knowledge* (TARK 98), pp. 43–56.

Baltag, A., Rafiee Rad, S. & Smets, S. (2019). Learning probabilities: Towards a logic of statistical learning. In *Proceedings of the seventeenth conference on theoretical aspects of rationality and knowledge* (TARK 17), EPTCS 297, pp. 35–49.

Baltag, A., & Renne, B. (2016). Dynamic epistemic logic. In *Stanford encyclopedia of philosophy*.

Baltag, A. & Smets, S. (2008a). The logic of conditional doxastic actions. In *Texts in logic and games, special issue on new perspectives on games and interaction* (Vol. 4, pp. 9–31), Amsterdam University Press.

Baltag, A., & Smets, S. (2008b). A qualitative theory of dynamic interactive belief revision. In *Texts in logic and games* (Vol. 3, pp. 9–58), Amsterdam University Press.

Baltag, A., & Smets, S. (2008c). Probabilistic dynamic belief revision. *Synthese, 165*(2), 179–202.

Board, O. (2004). Dynamic interactive epistemology. *Games and Economic Behavior*, *49*, 49–80.

Booth, R., & Meyer, T. (2006). Admissible and restrained revision. *Journal of Artificial Intelligence Research*, *26*, 127–151.

Bradley, R., & Drechsler, M. (2014). Types of uncertainty. *Erkenntnis*, *79*, 1225–1248.

Bradley, S., & Steele, K. (2014). Uncertainty, learning and the 'problem' of dilation. *Erkenntnis*, *79*, 1287–1303.

Cerreia-Vioglio, S., Maccheroni, F., Marinacci, M., & Montrucchio, L. (2013). Ambiguity and robust statistics. *Journal of Economics Theory*, *148*, 974–1049.

Chandler, J. (2014). Subjective probabilities need not be sharp. *Erkenntnis*, *79*, 1273–1286.

Darwiche, A., & Pearl, J. (1997). On the logic of iterated belief revision. *Artificial Intelligence*, *89*(1–2), 1–29.

Denoeux, T. (2000). Modeling vague beliefs using fuzzy-valued belief structures. *Fuzzy Sets and Systems*, *116*(2), 167–199.

Doob, J. L. (1971). What is a martingale? *American Mathematical Monthly*, *78*, 451–462.

Earman, J. (1992). *Bayes or bust: A critical examination of bayesian confirmation theory*, MIT press.

Edwards, W., Lindman, R., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Elkin, L., & Wheeler, G. (2018). Resolving peer disagreements through imprecise probabilities. *Nous*, *52*(2), 260–278.

Gaifman, H. (2016). A theory of higher order probabilities. In H. Arlo-Costa, V. F. Hendricks, & J. van Benthem (Eds.), *Readings in formal epistemiology* (Vol. 1, pp. 91–106), Springer.

Gaifman, H., & Snir, M. (1982). Probabilities over rich languages. *Journal of Symbolic Logic*, *47*, 495–548.

Gilboa, I., & Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, *18*, 141–153.

Gold, E. M. (1967). Language identification in the limit. *Information and control*, *10*(5), 447–474.

Grove, A. (1988). Two modellings for theory change. *Journal of Philosophical Logic*, *17*(2), 157–170.

Hájek, A. (2019). Interpretations of probability. In *The Stanford encyclopedia of philosophy*.

Hajek, A., & Smithson, M. (2012). Rationality and indeterminate probabilities. *Synthese*, *187*, 33–48.

Halpern, J. Y. (2003). *Reasoning about uncertainty*. Cambridge, MA: MIT Press.

Huber, P. J., & Strassen, V. (1973). Minimax test and Neyman–Pearson lemma for capacities. *The Annals of Statistics*, *1*, 251–263.

Hunter, J. K. (2012). *An introduction to real analysis*https://www.math.ucdavis.edu/~hunter/m125a/intro_analysis.pdf.

Huntley, N., Hable, R., & Troffaes, M. (2014). Decision making. In T. Augustin, F. P. A. Coolen, G. de Cooman, & M. C. M. Troffaes (Eds.), *Introduction to imprecise probabilities*.

Kelly, K. T. (1998). The learning power of belief revision. In *TARK'98: Proceedings of the 7th conference on theoretical aspects of rationality and knowledge* (pp. 111–124), Morgan Kaufmann Publishers Inc.

Kelly, K. T. (2008). *Ockham's razor, truth, and information. Handbook of the philosophy of information, Dordrecht*. Amsterdam: Elsevier.

Kelly, K. T. (2014). A computational learning semantics for inductive empirical knowledge. In A. Baltag & S. Smets (Eds.), *Outstanding contributions to logic: Johan van Benthem on logic and information dynamics* (Vol. 5, pp. 289–338), Springer.

Kelly, K. T., Schulte, O., & Hendricks, V. (1995). Reliable belief revision. In *Proceedings of the 10th international congress of logic, methodology, and philosophy of science* (pp. 383–398), Kluwer Academic Publishers.

Klibanoff, P., Marinacci, M., & Mukerji, S. (2005). A smooth model of decision making under ambiguity. *Econometrica*, *73*, 1849–1892.

Konieczny, S., & Perez, R. P. (2000). A framework for iterated revision. *Journal of Applied Non-Classical Logics*, *10*(3–4), 339–367.

Kuipers, T. A. F. (Ed.). (1987). *What is closer-to-the-truth? A parade of approaches to truthlikeness, poznan studies in the philosophy of the sciences and the humanities* (Vol. 10). Amsterdam: Rodopi.

Lehrer, K. (1990). *Theory of knowledge*. London: Routledge.

Leitgeb, H. (2017). *The stability of belief. How rational belief coheres with probability*. Oxford: Oxford University Press.

Levi, I. (1985). Imprecision and indeterminacy in probability judgment. *Philosophy of Science*, *52*, 390–409.

Lewis, D. (2000). *Counterfactuals*, Wiley, 1st edn 1973, 2nd edn.

Mayo-Wilson, C., & Wheeler, G. (2016). Scoring imprecise credences: A mildly immodest proposal. *Philosophy and Phenomenological Research*, *93*(1), 55–78.

Miller, D. (1974). Popper's qualitative theory of verisimilitude. *The British Journal for the Philosophy of Science*, *25*, 166–177.

Nayak, A. C. (1994). Iterated belief change based on epistemic entrenchment. *Erkenntnis*, *41*, 353–390.

Niiniluoto, I. (1987). *Truthlikeness*. Dordrecht: Reidel.

Niiniluoto, I. (1998). Verisimilitude: The third period. *The British Journal for the Philosophy of Science*, *49*(1), 1–29.

Paris, J. B. (1994). *The uncertain reasoner's companion: A mathematical perspective*. Cambridge: Cambridge Univ Press.

Paris, J. B. (2014). What you see is what you get. *Entropy*, *16*, 6186–6194.

Paris, J. & Rad, S. R. (2008). Inference processes for quantified predicate knowledge. In W. Hodges & R. de Queiroz (Eds.), *Logic, language, information and computation*, Springer LNAI (Vol. 5110, pp. 249–259).

Paris, J. & Rad, S. R. (2010). A note on the least informative model of a theory. In F. Ferreira, B. Lowe, E. Mayordomo, & L. Mendes Gomes (Eds.), *Programs, proofs, processes , CiE 2010*, Springer LNCS (Vol. 6158, pp. 342–351).

Paris, J. B., & Vencovska, A. (1997). In defence of the maximum entropy inference process. *International Journal of Approximate Reasoning*, *17*(1), 77–103.

Popper, K. R. (1976). A note on verisimilitude. *The British Journal for the Philosophy of Science*, *27*(2), 147–159.

Rad, S. R. (2017). Equivocation axiom for first order languages. In *Studia Logica, 105*(21).

Romeijn, J.-W., & Roy, O. (2014). Radical uncertainty: Beyond probabilistic models of belief. *Erkenntnis*, *79*(6), 1221–1223.

Rott, H. (2004). Stability, strength and sensitivity: Converting belief into knowledge. *Erkenntnis*, *61*(2–3), 469–493.

Rudin, W. (1953). *Principles of mathematical analysis*, McGraw-Hill Inc.

Savage, L. J. (1954). *Foundations of statistics*. New York: Wiley.

Schmeidler, D. (1986). Integral representation without additivity. *Proceedings of the American Mathematical Society, 97*(2).

Schmeidler, D. (1989). Subjective probability and expected utility without additivity. *Econometrica*, *57*(3), 571–587.

Seidenfeld, T. (2004). A contrast between two decision rules for use with (convex) sets of probabilities: Gamma-maximin versus E-admissibility. *Synthese*, *140*, 69–88.

Seidenfeld, T., Schervish, M. J., & Kadane, J. B. (2010). Coherent choice functions under uncertainty. *Synthese*, *172*, 157–176.

Skyrms, B. (2011). Resiliency, propensities, and causal necessity. In A. Eagle (Ed.), *Philosophy of probability: Contemporary readings*, Routledge.

Spohn, W. (2016). A survey of ranking theory. In H. Arlo-Costa, V. F. Hendricks, & J. van Benthem (Eds.), *Readings in formal epistemiology* (Vol. 1, pp. 303–350), Springer.

Stalnaker, R. (1996). Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy*, *12*, 133–163.

Tichy, P. (1974). On Popper's definitions of verisimilitude. *The British Journal for the Philosophy of Science*, *25*(2), 155–160.
Troffaesin, C. M. (2007). Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, *45*, 17–29.
van Benthem, J. (2007). Dynamic logic of belief revision. *Journal for Applied Non-Classical Logics*, *17*(2), 129–155.
van Benthem, J. (2011). *Logical dynamics of information and interaction*. Cambridge: Cambridge University Press.
van Ditmarsch, H., van der Hoek, W., & Kooi, B. (2007). *Dynamic epistemic logic*. Dordrecht: Springer.
Walley, P. (1996). Inferences from multinomal data: Learning about a bag of marbles. *Journal of the Royal Statistical Society Series B*, *58*, 3–57.
Walley, P. (2000). Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, *24*(2), 125–148.
Williamson, J. (2008). Objective Bayesian probabilistic logic. *Journal of Algorithms in Cognition, Informatics and Logic*, *63*, 167–183.
Williamson, J. (2010). *In defence of objective bayesianism*. Oxford: Oxford University Press.
Williamson, J. (2013). From Bayesian epistemology to inductive logic. *Journal of Applied Logic*, 2.
Williams, J., & Robert, G. (2014). Decision-making under indeterminacy. *Philosophers' Impreint*, *14*, 1–34.

Affiliations

Alexandru Baltag[1] · Soroush Rafiee Rad[1,2] · Sonja Smets[1,3]

✉ Sonja Smets
S.J.L.Smets@uva.nl

Alexandru Baltag
TheAlexandruBaltag@gmail.com

Soroush Rafiee Rad
Soroush.R.Rad@gmail.com

[1] Institute for Logic, Language and Computation (ILLC), University of Amsterdam, Amsterdam, The Netherlands

[2] Dutch Institute for Emergent Phenomena (DIEP), University of Amsterdam, Amsterdam, The Netherlands

[3] Department of Information Science and Media Studies, University of Bergen, Bergen, Norway