Learning Probabilities: Towards a Logic of Statistical Learning

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

[1] ILLC, University of Amsterdam, the Netherlands
{a.baltag,s.j.l.smets}@uva.nl
[2] University of Bayreuth, Germany
soroush.r.rad@gmail.com

Abstract. We propose a new model for forming beliefs and learning about unknown probabilities (such as the probability of picking a red marble from a bag with an unknown distribution of colored marbles). The most widespread model for such situations of 'radical uncertainty' is in terms of imprecise probabilities, i.e. representing the agent's knowledge as a set of probability measures. We add to this model a plausibility map, associating to each measure a plausibility number, as a way to go beyond what is known with certainty and represent the agent's beliefs about probability. There are a number of standard examples: Shannon Entropy, Center of Mass etc. We then consider learning of two types of information: (1) learning by repeated sampling from the unknown distribution (e.g. picking marbles from the bag); and (2) learning higher-order information about the distribution (in the shape of linear inequalities, e.g. we are told there are more red marbles than green marbles). The first changes only the plausibility map (via a 'plausibilistic' version of Bayes' Rule), but leaves the given set of measures unchanged; the second shrinks the set of measures, without changing their plausibility. Beliefs are defined as in Belief Revision Theory, in terms of truth in the most plausible worlds. But our belief change does not comply with standard AGM axioms, since the revision induced by (1) is of a non-AGM type. This is essential, as it allows our agents to learn the true probability: we prove that the beliefs obtained by repeated sampling converge almost surely to the correct belief (in the true probability). We end by sketching the contours of a dynamic doxastic logic for statistical learning.

Keywords: Radical uncertainty . Imprecise probabilities . Plausibility models . Statistical Learning . Belief Revision Theory

1 Introduction

Our goal in this paper is to propose a new model for *learning a probabilistic distribution*, in cases that are commonly characterized as those of "radical uncertainty" [22]. As an example, consider an urn full of marbles, coloured red, green and black, 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 probability distribution, she is typically left with a huge (usually infinite) *set* of probability assignments. If she never goes beyond what she knows, then her only 'rational' answer should be "I don't know": she in a state of *ambiguity*, and she should simply consider possible *all* distributions that are consistent with her background knowledge and observed evidence. Such a "Buridan's ass" type of rationality will not help our agent much in her 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. To form beliefs, the agent uses an AGM-type of *plausibility maximization*: she believes the most plausible distribution(s). So 'belief' is defined in our setting in the way that is standard in Logic and Belief Revision Theory: as "truth in all the most plausible worlds". The plausibility map encodes the agent's background knowledge and a priori assumptions about the world. For instance, an agent whose a priori assumptions include the Principle of Indifference will use Shannon entropy as her plausibility function, thus initially believing the most non-informative distribution(s). An agent who assumes some form of Ockham's Razor will use as plausibility some measure of simplicity, thus her initial belief will focus on the simplest distribution(s), etc. 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. probabilities distributions defined on the set of probability distributions). 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 the distributions with the *highest* plausibility play a role in defining 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 *linear inequality constraints* on that distribution. To see the differrence 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 biases of the coin. As such, any number of finite repeated samples *will not* shrink the set of possible biases, though they may make increase the plausibility of some biases. Thus this type of information changes only the plausibility map but leaves the given set of measures unchanged. The second type of information, on the other hand, shrinks the set of measures, without changing their plausibility. As 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 a 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 standards 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 "convergence in belief" is what we aim to capture in this paper.

Our belief revision mechanism after sampling is non-Bayesian (and also different from the AGM belief revision). It is essential that the agent does not keep only the 'prior' (i.e., the initially believed distribution), forgetting about the other possible distributions. Instead, after sampling she keeps all possibilities in store, but *revises her plausibility* map in the view of the new evidence, using a "plausibilistic analogue" of Bayes' Rule. 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.[1] Note also that the belief update induced by sampling does *not* satisfy all the standard AGM axioms. This is essential for learning the true probability from repeated sampling: since every sample is logically consistent with every distribution, an AGM learner would never change her initial belief!

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 former 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 (and as *update*, or "public announcement", in Logic), and satisfies all the standard AGM axioms.

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 different attitudes: the *set* of possible distributions (encoding the agent's *knowledge* about the correct distribution), and the *plausibility* map (encoding the agent's *beliefs*). The second type of (higher-order) information directly affects the agent's knowledge (by reducing

[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.

the set of possibilities), and only indirectly her beliefs (by restricting the plausibility map to the new set, so beliefs are only updated with fit the new knowledge). Dually, the first type of (sampling) evidence acts 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).
The plan of this paper follows. We start by reviewing some basic notions, results and examples on probability distributions (Section 2). Then in Section 3, we define our main setting (probabilistic plausibility frames), consider a number of standard examples (Shannon Entropy, Center of Mass etc), then formalize the updates induced by the two types of new information, and prove our main result on convergence-in-belief. In Section 4, we sketch the contours of a dynamic doxastic logic for statistical learning and in Section 5 we investigate unifying the two types of learning. We end with some concluding remarks and a brief comparison with other approaches to the same problem (Section 6).

2 Preliminaries and Notation

Take a finite set $O = \{o_1, \ldots, o_n\}$ and let $M_O = \{\mu \in [0,1]^O : \sum_{o \in O} \mu(o) = 1\}$ be the set of probability mass functions on O, which we identify with the corresponding probability functions on $\mathcal{P}(O)$. Let $\Omega = O^\infty = O^{\mathbb{N}}$ be the set of infinite sequences from O, which we shall refer to as *observation streams*. For any $\omega \in \Omega$ and $i \in \mathbb{N}$, we write ω_i for the i-th component of ω, and ω^i for its initial segment of length i, that is $\omega_1, \ldots, \omega_i$. For each $o \in O$ we define the sets o^j to be the cylinders $o^j = \{\omega \in \Omega;\ \omega_j = o\} \subseteq \Omega$. Let $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ be the σ-algebra of subsets of Ω generated by the cylinders. Every probability distribution $\mu \in M_O$ induces a unique probability function, $\hat{\mu}$ over $(\Omega, \mathcal{A})$ by setting $\hat{\mu}(o^j) = \mu(o)$ which extends to all of $\mathcal{A}$ using independence. Let $\mathcal{E}$ be the subalgebra of $\mathcal{A}$ that is closed under complementation and *finite* unions and intersections of the cylinder sets. Then $\mathcal{E}$ will capture the set of events generated by finite sequences of observations.

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 has landed Heads (res. Tails). 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, blue and Green. Then the set M_O will be the set of different distribution of coloured marbles in the urn, Ω will be streams of* R, B *and* G *representing infinite draws from the urn, and R^j (res. B^j or G^j) will be the set of streams of draws in which the j-th draw is a Red (res. Blue or Green) marble.*

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 $\mathcal{D}_O := \{\boldsymbol{x} \in [0,1]^n \mid \sum x_i = 1\}$, then every $\mu \in M_O$ can be identified with the point $\boldsymbol{\mu} \in \mathcal{D}_O \subset [0,1]^n$. Thus probability functions in $\mathcal{M}_O$ live in the space $\mathbb{R}^n$ (or more precisely $[0,1]^n$). In the other direction every $\boldsymbol{x} \in \mathcal{D}_O$ defines a probability function x on O by setting $x(o_i) = \boldsymbol{x}_i$. 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 $\mathbb{R}^n$,induced by the Euclidean metric: for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, put $d(\boldsymbol{x}, \boldsymbol{y}) := \sum_{i=1}^n \sqrt{(x_i - y_i)^2}$; a basis for the standard topology is given by the family of all *open balls* $\mathcal{B}_\epsilon(\boldsymbol{x})$ *centered at some point* $\boldsymbol{x} \in \mathbb{R}^n$ *with radius* $\epsilon > 0$; where

$$\mathcal{B}_\epsilon(\boldsymbol{x}) = \{\boldsymbol{y} \in \mathbb{R}^n : d(\boldsymbol{x}, \boldsymbol{y}) < \epsilon\}.$$

Proposition 1. *For a finite set O, the set of probability mass functions on O, $\mathcal{M}_O$, is compact in the standard topology.*

Proof. Check that the set $\{\boldsymbol{X} \in [0,1]^n \mid \sum_{i=1}^n x_i = 1\}$ is compact in $\mathbb{R}^n$.

We will make use of the following well known facts:

Proposition 2. *Let X, Y be compact topological spaces, $Z \subseteq X$ and $f : X \subseteq Y$*
(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 [10], Theorem 1.40 and Proposition 1.41.

Proposition 3. *Let X be a compact topological space and $f : X \to \mathbb{R}$ a continuous function on X. Then f is bounded and attains its supremum.*

Proof. See [10], 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 [29].

3 Probabilistic Plausibility Frames

A *probabilistic plausibility frame* over a finite set O is a structure $\mathcal{F} = (M, pla)$ where M is a subset of M_O, called the set of "possible worlds", and $pla : M_O \to [0, \infty)$ is a continuous function s.t. (1) the derivative pla' is also continuous, and (2) $pla(\mu) = 0$ implies $\mu(o) = 0$ for some $o \in O$.
So our possible worlds are just mass functions on O. Here are some canonical examples of probabilistic plausibility frames:

- (a) *Shannon Entropy* as plausibility: Let $pla : M_O \to [0, \infty)$ be given by the Shannon Entropy, $pla(\mu) = Ent(\mu) = -\sum_{o \in O} \mu(o) \log(\mu(o))$.
  Then (M_O, Ent) is a probabilistic plausibility frame. Here the most plausible distribution will be the one with highest Shannon entropy.
- (b) *Centre of Mass* as plausibility: Let $pla : M_O \to [0, \infty)$ be given by the Centre of Mass, $pla(\mu) = CM(\mu) = \sum_{o \in O} \log(\mu(o))$.
  Then (M_O, CM) is a probabilistic plausibility frame. Here the plausibility ranking will be given in terms of typicality, and higher plausibility will be given to those probability functions that are closer to the average of M_O.

Example 1. (continued). *In the absence of any information about the coin the set of possible biases will be the set M_O of* all *probability mass functions on $\{H, T\}$. Then (M_O, Ent) is a probabilistic plausibility frame, where the highest plausibility will be given to the distribution with highest entropy: the fair-coin distribution μ^{eq} (since for every $\nu \neq \mu^{eq}$ we have $Ent(\nu) < Ent(\mu^{eq})$).*

One of the main motivations for developing the setting that we investigate here is to capture the *learning process* as iterated revision that results from receiving new information. As was pointed out earlier one type of information essentially trims the space of possible probability measures by deleting certain candidates. There is however, a softer notion of revision, imposed by observations, that does not eliminate any candidate but rather changes the plausibility ordering over them. With this in mind, the next question we need to clarify is how the plausibility order is to be revised in light of new observations.

Definition 1 (*Conditionalization*). *Let $pla : M_O \to [0, \infty)$, and define $pla(.|.) : \mathcal{E} \times M_O \to [0, \infty)$, by $pla(\mu|e) := pla(\mu)\hat{\mu}(e)$. When $e \in \mathcal{E}$ is fixed, this yields a* conditional probability function *$pla_e : M_O \to [0, \infty)$ given by $pla_e(\mu) := pla(\mu \,|\, e)$.*

Conditionalising thus allows us to update the plausibility ranking of the set of probability distributions according to the new observations, and thus captures the notion *learning through sampling* pointed out above. The next three results in Lemma 1, and Propositions 5 and 4 ensure that the conditionalisation of the plausibility function given by Definition 1 behaves correctly. In particular, Lemma 1 and Corollary 4 show that the properties of a plausibility function in our frames is preserved by the conditionalisation and Proposition 5 guarantees that the result of repeated conditionalisation is independent of the order. This is important as it ensures that what the agents come to believe is the result of what they learn and not the order in which they learn them.

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

Proposition 4. *If pla is a plausibility function on M_O and $e \in \mathcal{E}$, then pla_e is a plausibility function.*

Proof. Follows from the definition using Lemma 1.

Proposition 5. *For $\mathcal{M}_O$ as above and $pla : \mathcal{M}_O \to [0,\infty)$ and $e, e' \in \mathcal{E}$: $(pla_e)_{e'} = pla_{e\cap e'}$.*

Proof. Let $\mu \in M$, then

$$(pla_{o^j})_{o'^k}(\mu) = pla_{o^j}(\mu)\hat{\mu}(o'^k) = pla(\mu)\hat{\mu}(o^j).\hat{\mu}(o'^k) = pla(\mu)\hat{\mu}(o^j \cap o'^k)$$

where the last equality follows from the independence assumption in *iid* case.

Example 1. (continued) *Take the frame (M_O, Ent) as before where M_O is the set of all biases of the coin and Ent is the Shannon Entropy. Remember that μ^{eq} is the unique maximiser of Ent on M_O. Let $e \in \mathcal{E}$, be the event that "the first three tosses of the coin have landed on Heads". After observing e, the new plausibility function is given by $pla_e(\mu) = pla(\mu)\hat{\mu}(e) = Ent(\mu)\hat{\mu}(e)$.*
Thus the most plausible probability function will no more be μ^{eq} and one with a bias towards Heads will become more plausible. Let μ_1, μ_2 and μ_3 be such that $\mu_1(Heads) = 3/4$, $\mu_2(Heads) = 0.8$ and $\mu_3(Heads) = 0.9$ then it is easy to check that $pla_e(\mu_1) < pla_e(\mu_2) > pla_e(\mu_3)$.

Our rule for updating plausibility relation weights the plausibility of each world with how much it respects the obtained evidence. In this way worlds that better correspond to the evidence are promoted in plausibility.

Proposition 6. *Let $M \subseteq M_O$ be closed. Then for all $e \in \mathcal{E}$, there exists some $\mu \in M$ with highest plausible (i.e. s.t. $pla_e(\mu) \geq pla_e(\mu')$ for all $\mu' \in M$).*

Proof. Using Lemma 1, the result follows as corollary of Proposition 3.

Definition 2 (Knowledge and Belief). *Let $P \subseteq M$ be a "proposition" (set of worlds) in a frame (M, pla). We say that P is* known*, and write $K(P)$, if all M-worlds are in P; i.e. $M \subseteq P$. We say that P is* believed *in frame $\mathcal{F} = (M, pla)$, and write $B(P)$, if and only if all "plausible enough" M-worlds are in P; i.e. $\{\nu \in M : pla(\nu) \geq pla(\mu)\} \subseteq P$ for some $\mu \in M$.*

Definition 3 (Two Forms of Conditionalization). *Let $P \subseteq M_O$ be a "proposition" (set of distributions). For an event $e \in \mathcal{E}$, we say that P is* believed conditional on *e in frame (M, pla), and write $B(P|e)$, if and only if all M-worlds that are "plausible enough given e" are in P; i.e. $\{\nu \in M : pla_e(\nu) \geq_e pla(\mu)\} \subseteq P$ for some $\mu \in M$. For a proposition $Q \subseteq M$, we say that P is* believed conditional on *Q in frame (M, pla), and write $B(P|Q)$, if and only if all plausible enough Q-worlds are in P; i.e. $\{\nu \in Q : pla(\nu) \geq pla(\mu)\} \subseteq P$ for some $\mu \in Q$.*

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 of M of all worlds represents the *tautological proposition* (corresponding to "no further higher-order information").
Belief is always consistent, and in fact it satisfies all the standard $KD45$ axioms of doxastic logic. 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 $B(P|Q)$). In fact, when the set of worlds is closed, our definition is equivalent to the standard definition of belief (and conditional belief) as "truth in all the most plausible worlds":

Proposition 7. *If $M \subseteq M_O$ is closed, then $B(P|e)$ holds if $\{\mu \in M : pla_e(\mu) \geq pla_e(\mu') \text{ for all } \mu' \in M\} \subseteq P$.*

Proof. Let $M \subseteq M_O$ be closed. Since pla_e is a continuous function, by Propositions 1, 2-1 and 3, there exists $\mu \in M$ such that for all $\mu' \neq \mu \in M$, $pla_e(\mu) \geq pla_e(\mu')$. Let $U_{pla_e} = \{\mu \in M \,|\, \forall \mu' \in M \; pla_e(\mu) \geq pla_e(\mu')\}$. Thus $U_{pla_e} \neq \emptyset$. Let $\mu \in U_{pla_e}$ and assume $U_{pla_e} \subseteq P$. Then we have $\{\nu \in M | pla_e(\nu) \geq_e pla_e(\mu)\} = U_{pla_e} \subseteq P$ and thus by definition $B(P|e)$.

We are now in the position to look into the learnability of the correct probability distribution via plausibility-revision induced by repeated sampling.

Theorem 2. *Take a finite set O of outcomes and consider a frame $\mathcal{M} = (M, pla)$ with $M \subseteq M_O$. Suppose that the correct probability is $\mu \in M$ and that $\mu(o_i) \neq 0$ for all i. Then, with μ-probability 1, the agent's belief will eventually stay arbitrarily close to the correct probability distribution after enough many observations. More precisely, for every $\epsilon > 0$, we have*

$$\mu(\{\omega \in \Omega \,|\, \exists K \; \forall m \geq K : \; B(\mathcal{B}_\epsilon(\mu) \,|\, \omega^m) \text{ holds in } M\}) = 1$$

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

To prove Theorem, we need a few well-known notions and facts:

Definition 4. *For $\mu \in M$, we define the set of μ-normal observations as the set of infinite sequences from O for which the limiting frequencies of each o_i correspond to $\mu(o_i)$ and we will denote this set by Ω_μ:*

$$\Omega_\mu = \{\omega \in \Omega \,|\, \forall o_i \in O \; \lim_{n \to \infty} \frac{|\{i \leq n \,|\, \omega_i = o_i\}|}{n} = \mu(o_i)\}.$$

Proposition 8. *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.

Lemma 2. *For $0 < p_1, \ldots, p_n < 1$ with $\sum p_i = 1$, the function $f(\boldsymbol{x}) = \Pi_{i=1}^n x_i^{p_i}$ on domain $\boldsymbol{x} \in \{\boldsymbol{z} \in (0,1)^n \mid \sum z_i = 1\}$ has $\boldsymbol{x} = \boldsymbol{p}$ as its unique maximizer on M_O.*

Proof. First we notice that $f(\boldsymbol{x}) \geq 0$ on $M_O = \{\boldsymbol{z} \in [0,1]^n \mid \sum z_i = 1\}$ and by Propositions 1 and 3 f has a maximum on M_O. For any point $\boldsymbol{z} \in M_O$ with any $z_i = 0$ (or $z_i = 1$) $f(\boldsymbol{z}) = 0$ thus f reaches its maximum on $\{\boldsymbol{z} \in (0,1)^n \mid \sum z_i = 1\}$.
To show the result, we will show that $\log(f(\boldsymbol{x}))$ has $\boldsymbol{x} = \boldsymbol{p}$ as its unique maximizer on this domain. The result then follows from noticing that $f(x) \geq 0$ and the monotonicity of log function on $\mathbb{R}^+$. To maximise $\log(f(\boldsymbol{x}))$ subject to condition $\sum_i x_i = 1$ we use Lagrange multiplier methods: let

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

Setting partial derivatives of G equal to zero we get,

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

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 $\boldsymbol{x}$ that maximises f should satisfy the condition $x_i = p_i$ and thus $\boldsymbol{p}$ is the unique maximiser for f.

Proof (Theorem 2). Since $\mu(\Omega_\mu) = 1$ (by the Strong Law of Large Numbers), it is enough to show that

$$\forall \epsilon > 0 \forall \omega \in \Omega_\mu \; \exists K \; \forall m \geq K : \; B(\{\nu | d(\mu, \nu) < \epsilon\} \,|\, \omega^m) \text{ holds in } M.$$

Let us fix some $\epsilon > 0$ and some $\omega \in \Omega_\mu$. We need to show that, there exists $\nu \in M$ such that for all large enough m, for any $\xi \in M$ if $pla(\xi \,|\, \omega^m) \geq pla(\nu \,|\, \omega^m)$, then $d(\xi, \mu) < \epsilon$. To show this, we will prove a stronger claim, namely that:

$$\exists K \; \forall m \geq K \; \forall \nu \in M_O \, (d(\nu, \mu) \geq \epsilon \Rightarrow pla(\mu \,|\, \omega^m) > pla(\nu \,|\, \omega^m)) \, .$$

(Note that the desired conclusion follows immediately from this claim: since we can then take μ itself to be the desired $\nu \in M$. Then by the above claim, no measures ξ in M_O with $d(\mu, \xi) \geq \epsilon$ satisfies $pla(\xi \,|\, \omega^m) \geq pla(\mu \,|\, \omega^m)$ and thus all measures, ν that satisfy this inequality have to satisfy $d(\mu, \nu) < \epsilon$.) By definition, for all $\nu \in M_O$ we have $pla(\nu \,|\, \omega^m) = pla(\nu) \cdot \hat{\nu}(\omega^m)$. By independence, we obtain that $pla(\nu \,|\, \omega^m) = pla(\nu) \cdot \Pi_{i=1}^n \nu(o_i)^{m_i} = pla(\nu) \cdot \Pi_{i=1}^n \nu_i^{m \cdot \alpha_{i,m}}$, were we have put $\nu_i := \nu(o_i)$ and $\alpha_{i,m} = \frac{m_i}{m}$, for all $1 \leq i \leq n$ and all $m \in N$. Note that, since $\omega \in \Omega_\mu$, we have that $\lim_{m \to \infty} \alpha_{i,m} = p_i$, for all $1 \leq i \leq n$, where we had put $p_i := \mu(o_i)$, for $1 \leq i \leq n$. In particular, for $\nu = \mu$ (so $\nu_i = \mu(o_i) = p_i$), we obtain that $pla(\mu \,|\, \omega^m) = pla(\mu) \cdot \Pi_{i=1}^n p_i^{m \cdot \alpha_{i,m}}$.
To prove the desired conclusion, it is enough (by the above representations of $pla(\nu \,|\, \omega^m)$ and $pla(\mu \,|\, \omega^m)$) to show that, for all big enough m and all $\nu \in M_O \setminus B_\epsilon(\mu)$, we have

$$pla(\nu) \cdot \Pi_{i=1}^n \nu_i^{m \cdot \alpha_{i,m}} < pla(\mu) \cdot \Pi_{i=1}^n p_i^{m \cdot \alpha_{i,m}} \tag{1}$$

Since $\lim_{m \to \infty} \alpha_{i,m} = p_i$, there must exist some N_1 such that $\frac{p_i}{2} \leq \alpha_{i,m} \leq 2 \cdot p_i$ for all $m \geq N_1$ and all $1 \leq i \leq n$. Let $\Delta = \{\nu \in M_O \,|\, \nu(o_i) = 0 \text{ for some } 1 \leq i \leq n\}$, and similarly for any $\delta > 0$, put $\Delta_\delta = \{\nu \in M_O \,|\, \nu(o_i) < \delta \text{ for some } 1 \leq i \leq n\}$, and so $\overline{\Delta_\delta} = \{\nu \in M_O \,|\, \nu(o_i) \leq \delta \text{ for some } 1 \leq i \leq n\}$ is its closure. Choose some $\delta > 0$ small enough such that we have $\Pi_{i=1}^n \nu_i^{2 \cdot p_i} < \Pi_{i=1}^n p_i^{\frac{p_i}{2}}$ for all $\nu \in \overline{\Delta_\delta}$ (-this is possible, since $\Pi_{i=1}^n \nu_i^{2 \cdot p_i} = 0 < \Pi_{i=1}^n p_i^{\frac{p_i}{2}}$ for all $\nu \in \Delta$, so the continuity of $\Pi_{i=1}^n \nu_i^{2 \cdot p_i}$ gives us the existence of δ). Hence, we have

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

Notice that by assumption $p_i = \mu(o_i) \neq 0$ for all $i = 1, \ldots, n$. The set $\overline{\Delta_\delta}$ is closed, hence the continuous functions $pla(\nu)$ and $\frac{\Pi_{i=1}^n \nu_i^{2 \cdot p_i}}{\Pi_{i=1}^n p_i^{\frac{p_i}{2}}}$ attain their supremum (maximum) on $\overline{\Delta_\delta}$. Let $K < \infty$ be the maximum of $pla(\nu)$, and $Q < 1$ be the maximum of $\frac{\Pi_{i=1}^n \nu_i^{2 \cdot p_i}}{\Pi_{i=1}^n p_i^{\frac{p_i}{2}}}$ on this set (-the fact that $Q < 1$ follows from the inequality above). Then there exists some $N_2 > N_1$, s.t. we have $Q^m < \frac{pla(\mu)}{K}$ for all $m > N_2$. Hence, for all $\nu \in \Delta_\delta$, we have:
$pla(\nu) \cdot \Pi_{i=1}^n \nu_i^{m \cdot \alpha_{i,m}} \leq K \cdot \Pi_{i=1}^n \nu_i^{m \cdot 2 \cdot p_i} \leq K \cdot (Q \cdot \Pi_{i=1}^n p_i^{\frac{p_i}{2}})^m = K \cdot Q^m \cdot \Pi_{i=1}^n p_i^{m \cdot \frac{p_i}{2}} < K \cdot \frac{pla(\mu)}{K} \cdot \Pi_{i=1}^n p_i^{m \cdot \alpha_{i,m}} = pla(\mu) \cdot \Pi_{i=1}^n p_i^{m \cdot \alpha_{i,m}}$
So we proved that the inequality (1) holds on Δ_δ. It thus remains only to prove it for all $\nu \in M' := M_O - (B_\epsilon(\mu) \cup \Delta_\delta)$, where $B_\epsilon(\mu) = \{\nu \in M_O \mid d(\mu, \nu) < \epsilon\}$. For this, note that $M' := M_O - (B_\epsilon(\mu) \cup \Delta_\delta)$ is closed and that $\nu_i \neq 0$ over this set (for all i) and thus by definition $pla(\nu) \neq 0$. Hence, (1) is equivalent over this set with:

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

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

$$\log(pla(\mu)) - \log(pla(\nu)) + \sum_{i=1}^n m \cdot \alpha_{i,m} \cdot (\log p_i - \log \nu_i) > 0. \tag{3}$$

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

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

Recall that $\alpha_{i,m} \geq \frac{p_i}{2}$ 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 $f(\nu) = 2 \cdot (\log(pla(\nu)) - \log(pla(\mu)))$ 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

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

(where at the end we used the fact, proved in Lemma 2, 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

$A = min_{M'}(g)$ on M'. But since g is non-zero on M', this minimum cannot be zero: $A = min_{M'}(g) \neq 0$. Similarly, since f is continuous and M' is closed, g is bounded and attains its supremum $B = max_{M'}(f) < \infty$ (which thus has to be finite). Take now some $N \geq max(N_2, \frac{B}{A})$. For all $m > N$, we have

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

for all $\nu \in M'$, as desired.

Corollary 1. *Suppose that $M \subseteq M_O$ is finite, and the correct probability is $\mu \in M$, with $\mu(o_i) \neq 0$ for all i. Then, with μ-probability 1, the agent's belief will settle on the correct probability distribution μ after finitely many observations:*

$$\mu(\{\omega \in \Omega \mid \exists K\ \forall m \geq K :\ B(\{\mu\} \mid \omega^m) \text{ holds in } M\}) = 1.$$

Proof. Apply the previous Theorem to some $\epsilon > 0$ small enough so that $\{\nu \mid d(\mu, \nu) < \epsilon\} \cap M = \{\mu\}$.

It is important to note the differences between our convergence result and the Savage style convergence results in the Bayesian literature that we mentioned in the Introduction. Savage's theorem only works for a finite set of hypotheses (corresponding to finite or countable M), so that the prior can assume a non-zero probability for each. Ours does not need this assumption and indeed, it works on the whole M_O, since we don't put a probability over hypothesis (probability measures), but rather a plausibility. Also, in the case of a finite set of hypotheses/distributions, our approach converges in finitely many steps (while Savage's still converges only in the limit).

4 Towards a Logic of Statistical Learning

In this section we will develop the logical setting that can capture the dynamics of learning described above. As was originally intended our logical language will be designed as to accommodate both type of information, i.e. finite observations and higher order information expressed in terms of linear inequalities. As we pointed out at the start there is a fundamental distinction between these two types of information which is reflected in the way that ingredients of our logical language are interpreted. The observations are interpreted in a σ-algebra $\mathcal{E} \subseteq \mathcal{P}(\Omega)$ and are not themselves formulas in our logical language as they do not correspond to properties over the set of probability measures. The reason, as described before, lies in the fact that no finite sequence of observations can rule out any possible probability distribution and as such do not single out any subset of the domain. The formulas of our logical language will instead be statements concerning the probabilities of observations given in terms of linear inequalities and logical combinations thereof as well as the statements concerning the dynamics arising from such finite observations.

Our set of *propositional variables* is the set of outcomes $O = \{o_1, \ldots, o_n\}$. The set of formulas, in our language, FL_{LS}, is inductively defined as

$$\phi ::= \top \mid \sum_{i=1}^{m} a_i w(o_i) \geq c \mid \phi \wedge \phi \mid \mid \neg\phi \mid K\phi \mid B(\phi|o) \mid B(\phi|\phi) \mid [o]\phi \mid [\phi]\phi$$

where $o_i \in O$, a_i's and c in $\mathbb{Q}$. The propositional connectives $\top, \neg, \wedge$ are standard. Letters K and B stand for knowledge and (conditional) belief operators, and $[o]$ and $[\phi]$ capture the *dynamics* of learning by an observation, o and by higher order information, ϕ respectively, and stand for "after observing o", and "after learning ϕ". Simple belief $B\phi$ is taken to be an abbreviation for $B(\phi|\top)$.

Definition 5 (Probabilistic Plausibility Models). *A* probabilistic plausibility model *over a finite set O is a structure $\mathcal{M} = (M, pla, v)$ where $M \subseteq M_O$, (M, pla) is a probabilistic plausibility frame and an evaluation function $v : O \to \mathcal{E}$ that assigns to each propositional variable o a cylinder set o^j.* [2]

Definition 6 (*Two types of update*). *Let $\mathcal{M} = (M, pla, v)$ be a probabilistic plausibility model, let $e \in \mathcal{E}$ be a sampling event, and let $P \subseteq M$ be a higher-order "proposition" (set of possible worlds, expressing some higher-order information about the world). The result of* updating the model with sampling evidence *e is the model $\mathcal{M}^e = (M, pla_e, v)$. In contrast, the result of* updating the model with proposition *P is the model $\mathcal{M}^P = (P, pla, v)$.*

Let $\mathcal{M} = (M, pla, v)$ be a probabilistic plausibility model. The semantics for formulas is given by inductively defining a satisfaction relation $\vDash$ between worlds and formulas. In the definition, we use the notation $\|\phi\|_{\mathcal{M}} := \{\mu \in M : \mathcal{M}, \mu \vDash \phi\}$:

$$\begin{aligned}
&\mathcal{M}, \mu \vDash \sum_{i=1}^{n} a_i w(o_i) \geq c && \iff \sum_{i=1}^{n} a_i \hat{\mu}(v(o_i)) \geq c \\
&\mathcal{M}, \mu \vDash \phi_1 \wedge \phi_2 && \iff \mathcal{M}, \mu \vDash \phi_1 \text{ and } \mathcal{M}, \mu \vDash \phi_2 \\
&\mathcal{M}, \mu \vDash \phi_1 \wedge \phi_2 && \iff \mathcal{M}, \mu \vDash \phi_1 \text{ or } \mathcal{M}, \mu \vDash \phi_2 \\
&\mathcal{M}, \mu \vDash \neg\phi && \iff \mathcal{M}, \mu \nvDash \phi \\
&\mathcal{M}, \mu \vDash K\phi && \iff \mathcal{M}, \nu \vDash \phi \text{ for all } \nu \in M \\
&\mathcal{M}, \mu \vDash B(\phi \,|\theta) && \iff B(\|\phi\|_{\mathcal{M}} \,|\, \|\theta\|_{\mathcal{M}}) \text{ holds in } (M, pla) \\
&\mathcal{M}, \mu \vDash B(\phi \,|\, o) && \iff B(\|\phi\|_{\mathcal{M}} \,|\, o) \text{ holds in } (M, pla) \\
&\mathcal{M}, \mu \vDash [o]\phi && \iff \mathcal{M}^o, \mu \vDash \phi \\
&\mathcal{M}, \mu \vDash [\theta]\phi && \iff \big(\mathcal{M}, \mu \vDash \theta \implies \mathcal{M}^\theta, \mu \vDash \phi\big)
\end{aligned}$$

As is standard, for a model $\mathcal{M} = (M, pla, v)$, let $\|\phi\|_{\mathcal{M}} = \{\nu \in M \mid \mathcal{M}, \nu \vDash \phi\}$ and we shall say that a formula ϕ is *valid in* $\mathcal{M}$ if and only if $\mathcal{M}, \mu \vDash \phi$ for all $\mu \in M$. Formula $\phi \in FL_{SL}$ is *valid* (in the logic L_{SL}) if it is valid in every model $\mathcal{M} = (M, pla, v)$.

[2] Notice that since we deal with i.i.d distributions the choice of j does not matter.

Proposition 9. *Let $\mathcal{M}$ be a probabilistic plausibility model. The set $B_{\mathcal{M}} = \{\phi \in FL_{PU} \mid \mathcal{M} \vDash B\phi\}$ is consistent.*

Proof. Take a probabilistic plausibility model $\mathcal{M} = (M, pla, v)$. Let $\phi \in B_{\mathcal{M}}$. We show that for any $\xi \in M$ there is some member of M that is at least as plausible as ξ but does not belong to $\neg\phi$ and thus by definition $\neg\phi \notin B_{\mathcal{M}}$.
Since $\phi \in B_{\mathcal{M}}$, by definition there exits $\mu \in M$ such that for all $\nu \in M$ with $pla(\nu) \geq pla(\mu)$, $\nu \in \|\phi\|$. Then if $pla(\xi) \geq pla(\mu)$, then $\xi \in \|\phi\|$ and thus $\xi \notin M \setminus \|\phi\| = \|\neg\phi\|$. Thus there exists some elements of M, namely, ξ itself that is at least as plausible of ξ but does not belong to $\|\neg\phi\|$. If $pla(\zeta) < pla(\mu)$ and since $\mu \in \|\phi\|$, $\mu \notin M \setminus \|\phi\| = \|\neg\phi\|$. Then again there is some member of M, namely μ that is more plausible than ξ but does not belong to $\|\neg\phi\|$.

Proposition 10. *Let $o \in O$ and $\phi, \theta, \xi \in FL_{SL}$. Then the following are valid formulas in L_{SL}*

- $w(o) \geq 0$
- $\sum_{o \in O} w(o) = 1$
- $K(\phi \to \theta) \to (K\phi \to K\theta)$
- $K\phi \to \phi$
- $K\phi \to KK\phi$
- $\neg K\phi \to K\neg K\phi$
- $B(\phi \to \theta) \to (B\phi \to B\theta)$
- $K\phi \to B\phi$
- $B\phi \to BB\phi$
- $\neg B\phi \to B\neg B\phi$

Proof. Notice that at each model $\mathcal{M}$ and each world μ, w is interpreted as a probability mass function, namely μ itself. The rest follow easily from the definition.

The dynamic operator in our logic that correspond to learning of higher order information, $[\phi]$, is essentially an AGM type update and satisfies the corresponding axioms, that is:

Proposition 11. *Let $\phi, \theta, \xi \in FL_{SL}$. Then the following are valid formulas in L_{SL}*

- $B(\phi \mid \phi)$
- $B(\theta \mid \phi) \to (B(\xi \mid \phi \wedge \theta) \leftrightarrow B(\xi \mid \phi))$
- $\neg B(\neg\theta \mid \phi) \to (B(\xi \mid \phi \wedge \theta) \leftrightarrow B(\theta \to \xi \mid \phi))$
- *If $\phi \leftrightarrow \theta$ is valid in $\mathcal{M}$ then so is $B(\xi \mid \phi) \leftrightarrow B(\xi \mid \theta)$.*

Proof. Notice that the plausibility function induces a complete pre-order on the set of worlds. The validity of the above formulas over such frames as well as the correspondence between these formulas and the AGM axioms are given by Board in [1].

Finally, we give without proofs some validities regarding the interaction of the dynamic modalities with knowledge modality and (conditional) belief.

Proposition 12. *Let $\phi, \theta, \xi \in FL_{SL}$. Then the following are valid formulas in L_{SL}*

- $[\phi]q \leftrightarrow (\phi \to q)$ *for atomic* q
- $[o]q \leftrightarrow\!\!\to q$ *for atomic* q
- $[\phi]\neg\theta \leftrightarrow (\phi \to \neg[\phi]\theta)$
- $[o]\neg\theta \leftrightarrow (\neg[o]\theta)$
- $[\phi](\theta \wedge \xi) \leftrightarrow ([\phi]\theta \wedge [\phi]\xi)$
- $[o](\theta \wedge \xi) \leftrightarrow ([o]\theta \wedge [o]\xi)$
- $[\phi]K\theta \leftrightarrow (\phi \to K[\phi]\theta)$
- $[o]K\phi \iff K[o]\phi$
- $[\phi]B(\theta \,|\, \xi) \iff (\phi \to B([\phi]\theta \,|\, \phi \wedge [\phi]\xi))$
- $[o]B(\phi \,|\, o') \iff B([o]\phi \,|\, o, o')$

5 Unifying the Two Types of Learning

In this section we will extend the setting in Section **??** and unify the two types of learning. We will generalise our setting in two ways. First we extend our mathematical setting in a way that allows us to treat observation and propositions on par and hence unify the two types of learning that have been treated separately so far. Second, we extend our language to be able to express observations in the language. This increases the expressibility of the language by allowing the linear inequalities to express information concerning more complex events rather than only single observations.
As before take a set (of outcomes) O, $\Omega = O^\infty$ and let $\mathcal{E} \subset \mathcal{P}(\Omega)$ be a σ-algebra generated by the set of cylinders $o_i^j = \{\omega \in \Omega \,|\, \omega_j = o_i\}$ and let M_O be a set of probability mass function on O.

Definition 7. *(Generalised Frame) A generalised probabilistic plausibility frame over O is a structure $\mathcal{F}_O = (\Sigma, pla)$ where $\Sigma \subseteq M_O \times \Omega$ and $pla : M_O \to [0, \infty)$ is a continuous function, such that $pla(\mu) = 0$ if and only if $\mu(o) = 1$ for some $o \in O$. The elements $\sigma = (\mu_\sigma, \omega_\sigma) \in \Sigma$ are called possible worlds.*

Example 3 *Let $O = \{H, T\}$, Ω be the set of infinite sequences from O, $\mathcal{E}$ a σ-algebra of subsets of Ω and $Ent : M_O \to [0, \infty)$ the Shannon Entropy function with respect to the partition $\{H, T\}$. Then $(M_O \times \Omega, Ent)$ is a generalised probabilistic plausibility frame. A possible world is a pair consisting of a probability function over O and an infinite sequence of H and T.*

Definition 8. *Let $A \subseteq \Sigma$ and $A^\mu = \{o \in \Omega \,|\, (\mu, o) \in A\}$. Define*

$$\mathcal{A}_\mathcal{E} = \{A \subseteq \Sigma \,|\, \forall \mu \in M \;\; (A^\mu \in \mathcal{E})\}$$

as the set of subsets of Σ that are globally measurable in M.

Proposition 13. *$\mathcal{A}_\mathcal{E}$ is a σ-algebra.*

Proof.

- Σ and $\emptyset$ are obviously in $\mathcal{A}_\mathcal{E}$.
- Let $A \in \mathcal{A}_\mathcal{E}$ and $A^c = \Sigma - A$. For all $\mu \in M_O$, $(\mu, o) \in \Sigma$ for all $o \in \Omega$, so $A^\mu \cup (A^c)^\mu = \Omega$ and $A^\mu \in \mathcal{E}$ and since $\mathcal{E}$ is a σ-algebra $(A^c)^\mu \in \mathcal{E}$ and so $A^c \in \mathcal{A}_\mathcal{E}$.
- Let $A_1, \ldots, A_n \in \mathcal{A}_\mathcal{E}$ then $A_1^\mu, \ldots, A_n^\mu \in \mathcal{A}_\mathcal{E}$. Thus $(A_1 \cup \ldots, \cup A_n)^\mu = A_1^\mu \cup \ldots \cup A_n^\mu \in \mathcal{E}$ and so $A_1 \cup \ldots, \cup A_n \in \mathcal{A}_\mathcal{E}$.

Given a generalised frame (Σ, pla), we can define the conditional plausibility functions, $pla(.\,|\,.) : M_O \times \mathcal{E} \to [0, \infty)$, as before, by

$$pla(\mu \,|\, e) = pla(\mu)\hat{\mu}(e).$$

The plausibility function *pla* (and more generally the conditional plausibility function) impose a pre-order on the set of probability functions in M_O. In the previous setting this was a pre-order on the set of worlds and hence induced a notion of belief. To achieve the same in this setting the plausibility ranking need to be lifted from the set of probability functions to a ranking over the set of worlds Σ. With slight abuse of notation we will denote the lifting of the plausibility function from M_O to Σ also by *pla*.

Definition 9. *Let (Σ, pla) be a generalised probabilistic plausibility frame and $\sigma = (\mu_\sigma, \omega_\sigma) \in \Sigma$. Define $pla : \Sigma \to [0, \infty)$,*

$$pla(\sigma) = pla(\mu_\sigma),$$

and $pla(.\,|\,.) : \Sigma \times \mathcal{A}_\mathcal{E} \to [0, \infty)$,

$$pla(\sigma \,|\, A) = pla(\mu_\sigma)\hat{\mu}_\sigma(A^{\mu_\sigma})$$

where as before $\hat{\mu}$ is the unique extension of μ to $\mathcal{E} \subseteq \mathcal{P}(\Omega)$.

Example 4 1. (continued) *Let*

$$A = \Sigma_1 \cup \Sigma_2 \cup \Sigma_3,$$

where $\Sigma_1 = \{(\mu_1, \omega) \,|\, \omega_i = H, \omega_2 = T\}$ similarly let $\Sigma_2 = \{(\mu_1, \omega) \,|\, \omega_1 = T, \omega_2 = H\}$ and $\Sigma_3 = \{(\mu_2, \omega) \,|\, \omega_1 = H, \omega_2 = T, \omega_3 = H\}$. Then Then A_{μ_1}, for example, is the event that exactly one of the first two tosses lands Heads.

$$A^{\mu_1} = \{< H, T, \ldots >\} \cup \{< T, H, \ldots >\}, \text{and}$$

$$A^{\mu_2} = \{< H, T, H, \ldots >\}.$$

Next, let $\sigma_1 \in \Sigma_1$ and $\sigma_3 \in \Sigma_3$. Then

$$pla(\sigma_1 \,|\, A) = pla(\mu_1)\hat{\mu_1}(A^{\mu_1}).$$

Let μ_1 be the probability measure given by $\mu_1(Heads) = 0.8$ and μ_2 be the one for which $\mu_2(H) = 2\mu_2(T)$. Then before learning A, $pla(\sigma_1) < pla(\sigma_3)$ as the Shannon entropy of μ_2 is higher than μ_1. However, after learning A,

$$pla(\sigma_1 \,|\, A) = Ent(\mu_1).\hat{\mu_1}(A^{\mu_1}) \approx 0.71, and$$

$$pla(\sigma_3 \,|\, A) = Ent(\mu_2)\hat{\mu_2}(A^{\mu_2}) \approx 0.4$$

and we have $pla(\sigma_1 \,|\, A) > pla(\sigma_3 \,|\, A)$.

We can now consider an extended logical language L_{PU^*} that can express both the final observations and the higher order information given bt inequalities.

5.1 The Syntax of L_{PU^*}

Fix a finite set $O = \{o_1, \ldots, o_n\}$ of outcomes and let $Prop = \{o^j \,|\, o \in O, j \in \mathbb{N}\}$ be a set of atomic propositional variables. The set of formulas of L_{PU^*}, is inductively defined as

$$\phi ::= o^j \,|\, \top \,|\, \sum_{i=1}^{m} a_i w(\phi_i) \geq c \,|\, \phi_1 \wedge \phi_2 \,|\, \neg\phi \,|\, K\phi \,|\, B\phi \,|\, [\phi_1]\phi_2$$

where $o^j \in Prop$, a_i's and c in $\mathbb{Q}$, K and B stand for knowledge and belief operators and $[\phi]\psi$ stands for "after learning ϕ, ψ is true".

5.2 The Semantics of L_{SL^*}

Definition 10. *A generalised, probabilistic plausibility model $\mathcal{M}_O = (\Sigma, pla, v)$ where (Σ, pla) is a generalised probabilistic plausibility frame and $v : Prop \to \mathcal{E}$ a valuation of atomic propositions in $\mathcal{E}$ defined by*

$$v(o^j) = \{\sigma \in \Sigma \,|\, \omega_\sigma(j) = o\}.$$

Given a probabilistic plausibility model $\mathcal{M}_O = (\Sigma, pla, v)$ the semantics for FL_{SL} is given by the function $\|.\| : FL_{PU^*} \to \mathcal{A}_\mathcal{E}$ defined by,

$$\sigma \in \|\phi\| \iff \mathcal{M}, \sigma \vDash \phi$$

where

- $\mathcal{M}, \sigma \vDash o^j$ iff $\sigma \in M \times v(o^j)$,
- $\mathcal{M}, \sigma \vDash \sum_{i=1}^{n} a_i w(\psi_i) \geq c$ iff $\sigma \in \{\mu \in M \,|\, \sum_{i=1}^{n} a_i \hat{\mu}(\|\psi_i\|^\mu) \geq c\} \times \Omega$
- $\mathcal{M}, \sigma \vDash \phi_1 \wedge \phi_2$ iff $\mathcal{M}, \sigma \vDash \phi_1$ and $\mathcal{M}, \sigma \vDash \phi_2$
- $\mathcal{M}, \sigma \vDash \neg\phi$ iff $\mathcal{M}, \sigma \nvDash \phi$
- $\mathcal{M}, \sigma \vDash K\phi$ iff $\mathcal{M}, \delta \vDash \phi$ for all $\delta \in \Sigma$.
- $\mathcal{M}, \sigma \vDash B\phi$ if $\exists \delta \in \Sigma, \forall \gamma \in \Sigma s.t. pla(\gamma) \geq pla(\delta)$, $\mathcal{M}, \gamma \vDash \phi$.

– $\mathcal{M}, \sigma \vDash [\phi_1]\phi_2$ iff $\mathcal{M}, \sigma \vDash \phi_1$ and $\mathcal{M}^{\phi_1}, \sigma \vDash \phi_2$ where $\mathcal{M}^{\phi_1} = (\|\phi_1\|, pla(.\mid \|\phi_1\|), v)$.

Proposition 14. *The function $\|.\|$ is well defined, i.e, for all $\phi \in FL_{PU}$, $\|\phi\| \in \mathcal{A}_{\mathcal{E}}$.*

Proof. By induction on the ϕ.

– For $\phi = p$ a propositional variable, $\|p\| = M \times v(p)$. Thus for each $\mu \in M$, $\|p\|^{\mu} = v(p) \in \mathcal{E}$ thus, $\|p\| \in \mathcal{A}_{\mathcal{E}}$.
– For $\phi = \sum_{i=1}^{n} a_i w(\psi_i) \geq c$, $\|\phi\| = N \times \Omega$ for some $N \subseteq M$ then. Thus for each $\mu \in M$, $\|\phi\|^{\mu} = \Omega$ or $\|\phi\|^{\mu} = \emptyset$, thus, $\|\phi\| \in \mathcal{A}_{\mathcal{E}}$.
– For $\phi = \phi_1 \wedge \phi_2$, with $\|\phi_1\|, \|\phi_2\| \in \mathcal{A}_{\mathcal{E}}$. But $\mathcal{A}_{\mathcal{E}}$ is a σ-algebra thus $\|\phi_1\| \cap \|\phi_2\| \in \mathcal{A}_{\mathcal{E}}$.
– For $\phi = \neg\phi_1$, with $\|\phi_1\| \in \mathcal{A}_{\mathcal{E}}$ since $\mathcal{A}_{\mathcal{E}}$ is a σ-algebra we have $(\Sigma - \|\phi_1\|) \in \mathcal{A}_{\mathcal{E}}$.
– For $\phi = B\phi_1$, $\|\phi\| = \Sigma \in \mathcal{A}_{\mathcal{E}}$ or $\|\phi\| = \emptyset \in \mathcal{A}_{\mathcal{E}}$.
– For $\phi = [\phi_1]\phi_2$, with $\|\phi_1\|, \|\phi_2\| \in \mathcal{A}_{\mathcal{E}}$, $\forall \mu \in M$, $\|\phi_2\|^{\mu} \in \mathcal{A}_{\mathcal{E}}$, thus $\{\mu \mid (\mu, \omega) \in \|\phi_1\|\} \subseteq M$. Thus for all $\sigma \in \Sigma^{\phi_1} = \|\phi_1\|$, and all $\nu \in \{\mu \mid (\mu, \omega) \in \|\phi_1\|\}$ we have $\|\phi_2\|^{\nu} \in \mathcal{A}_{\mathcal{E}}$.

Proposition 15. *For a formula $\phi := \sum_{i=1}^{n} a_i w(\psi_i) \geq c$, $pla(.\mid \|\phi\|) = pla$.*

Definition 11 (Regular Worlds). *In a model $\mathcal{M}_O = (\Sigma, pla, V)$, we shall call a world $\sigma = (\mu, \omega) \in \Sigma$* regular *if for all $o \in O$*

$$\lim_{n \to \infty} \frac{|\{i \leq n \mid \omega(i) = o\}|}{n} = \mu(o),$$

we denote the set of regular worlds in Σ by $Reg(\Sigma)$.

Let Φ_{μ} and Φ_{μ}^{ϵ} be the propositions asserting that the correct probability function is μ and that the correct probability function is in an ϵ-neighbourhood of μ respectively.

$$\Phi_{\mu} = \{(\nu, \omega) \in \Sigma \mid \nu = \mu\}, and$$
$$\Phi_{\mu}^{\epsilon} = \{(\nu, \omega) \in \Sigma \mid d(\nu, \mu) \leq \epsilon\}.$$

We can now give an analogous of Proposition 8 that is a slight variation of the strong law of large numbers.

Proposition 16. *For every probability function μ,*

$$\mu(Reg(\Sigma) \cap \Phi_{\mu}) = 1$$

that is, almost all worlds in Σ are regular.

We are now in the position to state the analogous of the learning results of Theorem 2 for the extended setting.

Theorem 3. *Consider a model $\mathcal{M}_O = (\Sigma, pla, v)$ with $O = \{o_1, \ldots, o_n\}$, $M \subseteq M_O$ and $\Sigma = M \times \Omega$.*

- *(i) If M is finite then*

$$\mu(\{\sigma \in \Phi_\mu \,|\, \exists N\ \forall m \geq N \quad B(\Phi_\mu \,|\, \omega_\sigma^m)\}) = 1.$$

- *(ii) If M is infinite then*

$$\mu(\{\sigma \in \phi_\mu \,|\, \forall \epsilon > 0\ \exists N_\epsilon\ \forall m \geq N_\epsilon \quad B(\Phi_\mu^\epsilon \,|\, \omega_\sigma^m)\}) = 1.$$

Proof. Using Proposition 16, to show (i), it would be enough to show that for every regular $\sigma \in \phi_\mu$, there exists N such that for all $m \geq N$, $[\omega_\sigma(1), \ldots, \omega_\sigma(m)]B(\phi_\mu)$. Notice that for a regular $\sigma = (\mu_\sigma, \omega_\sigma)$, ω_σ is μ_σ-normal. The proof then follows by slight modification of the proof of Theorem 2.

Definition 12. *Let $M = (\Sigma, pla)$ and let $c \in (0, 1)$,*

- *$\Delta \subset \Sigma$ is called convex if for every $\sigma = (\mu_\sigma, \omega_\sigma), \delta = (\mu_\delta, \omega_\delta) \in \Delta$ and $\nu = c\mu_\sigma + (1 - c)\mu_\delta$*

$$(\nu, \omega) \in \Sigma \implies (\nu, \omega) \in \Delta.$$

- *A sentence ϕ is called convex if $||\phi||$ is a convex subset of Σ.*

Proposition 17. *Let $M = (\Sigma, pla, v)$ be a probabilistic plausibility model. Then every $\phi \in FL^+$ is convex.*

Proof. By induction on the structure of the formula.

- $\phi := p$. Then $\|\phi\| = M \times v(p)$.
- $\phi := \theta \wedge \psi$ with ψ, θ convex. Then $\|\phi\| = \|\psi\| \cap \|\theta\|$ and the intersection of two convex sets is convex
- $\phi := \sum_{i=1}^n a_i w(\phi_i) \geq c$. Then if $\sigma_1, \sigma_2 \in \|\phi\|$ we have $\sum_{i=1}^n a_i \mu_{\sigma_1}(\phi_i) \geq c$ and $\sum_{i=1}^n a_i \mu_{\sigma_2}(\phi_i) \geq c$ then for every $d \in (0\,,\, 1)$ and $\mu = d\mu_{\sigma_1} + (1 - d)\mu_{\sigma_2})$, $\sum_{i=1}^n a_i w(\phi_i) \geq c$ and so for every $\sigma = (\mu, \omega) \in \Sigma$, $\sigma \in \|\phi\|$ and so ϕ is convex.
- $\phi := K\psi$. Then $\|\phi\| = \Sigma$ or $\|\phi\| = \emptyset$.
- $\phi := B^\theta \psi$ with θ convex. Then $\|\phi\| = \Sigma$ or $\|\phi\| = \emptyset$.
- $\phi := [\theta]\psi$. Let $\sigma = (\mu_\sigma, \omega_\sigma), \delta = (\mu_\delta, \omega_\delta) \in \|[\theta]\psi\| \subseteq \Sigma$. Then $M, \sigma \vDash \theta$ and $M, \delta \vDash \theta$. For $d \in [0,\, 1]$ let $\mu = d\mu_\sigma + (1 - d)\mu_\delta)$ then since $\|\theta\|$ is convex

$$(\mu, \omega) \in \Sigma \implies (\mu, \omega) \in \|\theta\|$$

  Thus if Σ^θ be the set of state of M^θ then Σ^θ is convex. Then since $\|\psi\|$ is also convex in Σ^θ we have

$$\sigma, \delta \in \Sigma^\theta \cap \|\psi\| \to (\mu, \omega) \in \Sigma^\theta \cap \|\psi\|\,.$$

We will then hve the analogus of **??** for the extended setting as well.

Proposition 18. *Let $\mathcal{M} = (M, pla, v)$ be a generalised probabilistic plausibility model where pla is a convex function. Then for any $\phi \in FL^+$ we have*

$$\mathcal{M} \vDash B\left(\sum_{i=1}^m a_i w(o_i) \geq c \,|\, \phi\right) \vee B\left(\sum_{i=1}^m a_i w(o_i) < c \,|\, \phi\right)$$

6 Conclusion and Comparison with Other Work

We studied forming beliefs about unknown probabilities in the situations that are commonly described as the 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 [2, 4, 9, 12, 23–25] and on different approaches for decision making based on them [3, 11, 18, 26–28, 31, 32] 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 [14–17, 20, 21].

Another approach to deal with these scenarios in the Bayesian literature come from the series of convergence results collectively referred to as "washing out of the prior". The idea, which traces back to Savage, see [6, 30], 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.[3] 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. 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 a 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). We end by sketching the contours of a dynamic doxastic logic for statistical learning. Our belief operator satisfies

[3] To be more precise, if one starts with a prior probability for an event A, and keeps updating this probability by conditionalizing 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 has 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 [5]. There are also several generalisations of these results, e.g. Gaifman and Snir [8].

all the axioms of standard doxastic logic, and one form of conditional belief (with propositional information) satisfies the standard AGM axioms for belief revision. But the other form of conditioning (with sampling evidence) does not satisfy these axioms, and this is in fact essential for our convergence results.

References

1. Board, O., "Dynamic interactive epistemolog", in *Games and Economic Behavior*,49: 49–80, 2004.
2. Bradley, R., & Drechsler, M., " Types of Uncertainty", in *Erkenntnis*, 79: 1225–1248, 2014.
3. Bradley, S., & Steele, K., "Uncertainty, Learning and the 'Problem' of Dilation", in *Erkenntnis*, 79: 1287–1303, 2014.
4. Chandler, J., "Subjective Probabilities Need Not Be Sharp", in *Erkenntnis*, 79: 1273–1286, 2014.
5. Doob,J. L. "What Is a Martingale?" in *American Mathematical Monthly* 78:451-462, 1971.
6. Edwards, W., Lindman, R, and Savage, L. J., "Bayesian Statistical Inference for Psychological Research", in *Psychological Review* 70: 193-242, 1963.
7. Earman, J. " Bayes or Bust: A Critical Examination of Bayesian Confirmation theory", MIT press, 1992.
8. Gaifman, H., and Snir, M."Probabilities over Rich Languages", in *Journal of Symbolic Logic* 47:495-548, 1982.
9. Hajek, A., & Smithson, M., "Rationality and Indeterminate Probabilities", in *Synthese*, 187: 33–48, 2012.
10. Hunter, J. K., *An Introduction to Real Analysis*, available at *https* : *//www.math.ucdavis.edu/ hunter/m*125*a/intro*$_a$*nalysis.pdf*
11. Huntley, N., Hable, R., & and Troffaes, M., "Decision making", in Augustin et al, 2014.
12. Levi, I., "Imprecision and Indeterminacy in Probability Judgment", in *Philosophy of Science*, 52:390–409, 1985.
13. Paris, J. & Rafiee Rad, S., "Inference Processes for Quantified Predicate Knowledge", in *Logic, Language, Information and Computation, WoLLIC*, Eds. W. Hodges and R. de Queiroz, Springer LNAI, 5110: 249–259, 2008.
14. Paris, J. & Rafiee Rad, S., "A Note On The Least Informative Model of A Theory", in *Programs, Proofs, Processes , CiE 2010*, Eds. F. Ferreira, B. Lowe, E. Mayordomo, & L. Mendes Gomes, Springer LNCS, 6158: 342–351, 2010.
15. Paris, J.B. & Vencovska, "In defence of the maximum entropy inference process", in *International Journal of Approximate Reasoning*,17(1): 77-103, 1997.
16. Paris, J. B., "What you see is what you get?, in *Entropy*,16: 6186?6194, 2014. .
17. Rafiee Rad, S., "Equivocation Axiom for First Order Languages", in em Studia Logica, 105(21), 2017.
18. Troffaesin, C. M., "Decision making under uncertainty using imprecise probabilities", in *International Journal of Approximate Reasoning*, 45:17-29, 2007
19. Williamson, J., "From Bayesian epistemology to inductive logic?, in *Journal of Applied Logic*, 2, 2013.
20. Williamson, J., "Objective Bayesian probabilistic logic?, in *Journal of Algorithms in Cognition, Informatics and Logic*, 63:167-183, 2008.

21. Williamson, J., *In Defence of Objective Bayesianism*, Oxford University Press, 2010.
22. Walley, P. "Inferences from Multinomal Data: Learning about a bag of marbles", in *Journal of the Royal Statistical Society Series B*, 58:3-57, 1996.
23. Walley, P. "Towards a unified theory of imprecise probability", in *International Journal of Approximate Reasoning*, 24(2): 125-148, 2000.
24. Denoeux, T., "Modeling vague beliefs using fuzzy-valued belief structures", in *Fuzzy Sets and Systems*, 116(2):167-199, 2000.
25. Romeijn, J-W. & Roy, O., "Radical Uncertainty: Beyond Probabilistic Models of Belief", in *Erkenntnis*, 79(6):1221–1223, 2014.
26. Elkin, L. & Wheeler, G., "Resolving Peer Disagreements Through Imprecise Probabilities", in *Nous*, forthcoming.
27. Mayo-Wilson, C. & Wheeler, G. "Scoring Imprecise Credences: A Mildly Immodest Proposal" , in *Philosophy and Phenomenological Research*, 93(1): 55?78, 2016.
28. Seidenfeld, t., "A contrast between two decision rules for use with (convex) sets of probabilities: Gamma-maximin versus E-admissibility", in *Synthese*, 140:69–88, 2004.
29. Rudin, W., *Principles of Mathematical Analysis*, McGraw-Hill Inc, 1953.
30. Savage, L. J. *Foundations of Statistics*, New York: John Wiley, 1954.
31. Seidenfeld, T., Schervish, M. J., & Kadane, J. B., "Coherent choice functions under uncertainty", in *Synthese*, 172: 157–176, 2010.
32. williams, J. & Robert, G., "Decision-making under indeterminacy", in *Philosophers' Impreint*, 14:1–34, 2014.