Bridging statistical learning theory and dynamic epistemic logic

Paulius Skaisgiris Joel Saarinen

University of Amsterdam
Institute for Logic, Language and Computation
{paulius.skaisgiris, joel.saarinen}@student.uva.nl

5th June, 2024

Abstract

This paper establishes a connection between statistical learning theory (SLT) and formal learning theory (FLT). We use this insight to connect SLT and dynamic epistemic logic (DEL) models via the already established FLT and DEL bridge due to [Gierasimczuk, 2009]. Specifically, we demonstrate that the uniform convergence property of a hypothesis space implies the finite identifiability of its corresponding epistemic space which, in turn, can be modelled in DEL [Gierasimczuk, 2009]. This paper thus lays the foundations of an alternative way to introduce probabilistic reasoning into DEL, reason about statistical learning scenarios topologically, and investigate the epistemology of statistical learning theory.

Contents

1 Introduction and Motivation

Learning is a process greatly important to the navigation of the world around us. Specifically, in the broadest sense, learning for an agent may be thought of as the means by which they come to know new information. Learning may be done through observing something in the external world through sensory perception systems (for example, looking at the clock on the wall and learning that the time is noon), or by deducing a new claim from existing knowledge the agent holds in their heads (for example, if the agent knows the independent facts that "Mark will go to the park if it is sunny" and that "it is sunny", they can logically deduce and thus learn the fact that "Mark is at the park"), or through any method in between. Colloquially, these types of learning may be considered to be of a more *a posteriori* versus *a priori* nature, respectively, which is merely one way by which learning processes can differ.

Why exactly is learning important? One reason for this is that agents – whether human, or artificial – have goals, and achieving goals is easy to the extent that one has learned an accurate model of the world. For example, you might be investing in real estate, choosing between a few different options, and to settle upon the right choice (where the "right" choice refers to the one that maximizes some criteria you care about, such as net profit from selling it later) you presumably think about to what extent a given choice has maximized that criteria in the past (for example, by thinking about how much values of properties in a certain area have risen over time) and then choose the one with the highest expected reward with regards to that criteria.

In other words, you have *learned* a certain model of the world – a model about how different properties maximize your reward – from either investing in properties directly yourself or by observing the investments of others in the past. You then use that model to make a prediction which matters, in some way, to your well-being as an agent in the end. This means that learning matters to your well-being as an agent.

Having discussed the meaning of learning in more detail and motivated its importance, a natural follow-up question is what an optimal theory of learning might look like, in the prescriptive sense. Intuitively, as we've gestured at, a good theory should at least involve providing the learner with an accurate picture of the world. If the frameworks that form an agent's learning machinery cause the agent to possess beliefs about the world that do not reflect the actual world, and thus (for example) the agent's beliefs about what properties are good investments do not actually turn out to be good investments, it would be difficult to argue for this learning framework being good, insofar as this means enabling the agent to maximize reward. A good theory is conceivably also not rigid, in the sense that it should not lead a learner to be overconfident about its beliefs, and be open to changing them in the light of relevant evidence. All this being said, we postpone further

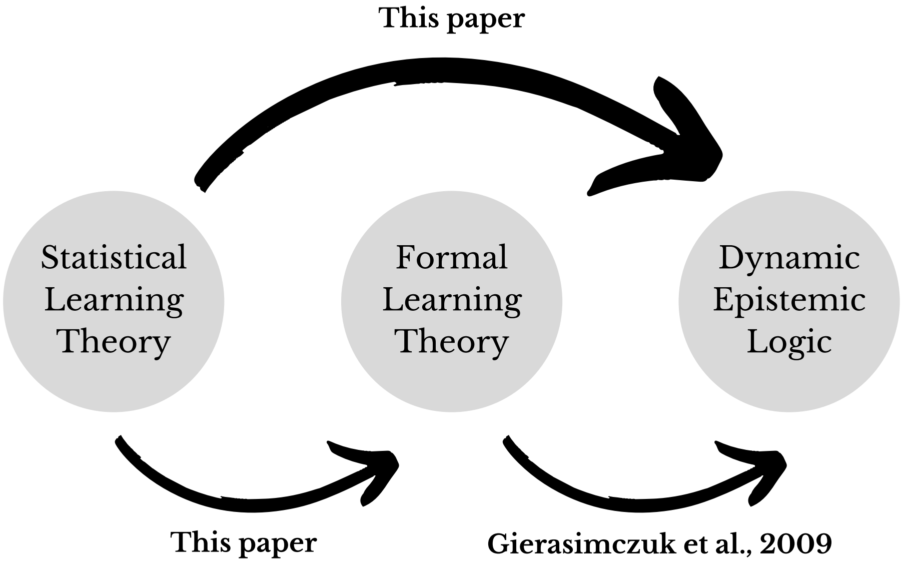

Figure 1: The high-level plan of this paper.

discussion of any sort of ideal theory, and claim only that a given theory may possess only one component of such an ideal. If this holds, combining, or bridging different theories of learning, would therefore conceivably get us closer to an ideal.

The motivation for working with these two frameworks specifically is as follows. These two frameworks exemplify a key way in which learning frameworks more generally can differ. Namely, some frameworks (such as DEL), when embedded into an agent, allow the agent to learn new information with certainty, while others (such as SLT), when embedded, supply the agent with information they do not necessarily have full confidence in. Intuitively, we might prefer to be more certain about how the world appears than uncertain – as having to update our models about the world takes time and resources – but being open-minded could reduce minimize risks that stem from placing too much credence in a model of the world that will eventually turn out to be wrong. Thus, there are advantages and disadvantages to both frameworks, which will soon be discussed in more length.

In this paper, we explore the prospect of bridging the aforementioned SLT and DEL learning frameworks with the motivation of creating a "fused" framework that combines the advantages of logic-based and probabilistic frameworks while minimizing the weaknesses. To do this, we will first explain SLT and DEL and the motivations for developing them in more detail, so as to set up the bridging. Then, we bridge the two frameworks by stating a correspondence between fundamental objects in both frameworks. We subsequently use this correspondence to show how uniform convergence of a hypothesis space can be modelled in FLT. We conclude by briefly considering what further work could be done to connect these frameworks.

2 Statistical Learning Theory

Statistical learning theory (SLT) was developed to serve as the mathematical foundations for machine learning. More precisely, some of the questions that SLT seeks to answer are:

1. How much training data is needed to minimize the risk of error to a desired extent with a certain likelihood? (probably approximately correct learning)
2. How can we describe the extent to which a set of data is learnable at all? (Vapnik-Chervonenkis-dimension)
3. How well is a class of possible predictor functions able to fit random noise? (Rademacher complexity)

SLT is used mainly in the context of supervised learning problems. That is, the theory of statistical learning deals with making predictions about regression and classification tasks, as opposed to ones where algorithms are trying to learn patterns from unlabeled data [Shalev-Shwartz and Ben-David, 2014]. Thus, in discussing SLT as a theory of learning, the type of learning we will be concerned with is supervised learning.

We can think of SLT as defining and modeling learning in the following manner. When a learner (implicitly, a computer program, specifically a learning algorithm) is said to have learned something, it means that, with the learner having seen multiple instances of ordered pairs $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, it is able to determine the correct label y for a new given vector of features $\mathbf{x}$, with this pair $(\mathbf{x}, y)$ not in the initial set of ordered pairs with an arbitrarily high degree of probability.

More formally, following the definitions of [Shalev-Shwartz and Ben-David, 2014], the learner is given access to a set $S = \{(\mathbf{x}_1, y_1), \ldots (\mathbf{x}_n, y_n)\}$ of training data, where $\mathbf{x} \in X$ and X is the set of objects that we want to label and $y \in Y$ where Y is the set of possible labels (usually assumed to be $\{0, 1\}$ in literature for simplicity), generated by some (unknown to the learner) probability distribution D. The task of the learner is to choose a classifier function $h : X \to Y$, from a class of hypotheses $\mathcal{H}$, that predicts the label of new domain points.

The measure of success for the learner is the *risk* of the classifier that the learner selects, or the probability that the classifier outputs the correct label for a given point. Risk may be defined in a few different ways, but perhaps one of the most canonical is empirical risk (with respect to a given set of training data S), denoted by:

$$L_S(h) := \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}$$

and interpreted as the proportion of the training data that h classifies the input differently from the (unknown to the learner) ideal labeling function f. Another common type of risk may simply give us the probability, with respect to the underlying distribution generating the training data, that h will differ from f. This may simply be expressed as:

$$L_{D,f}(h) = \mathbb{P}_{x \sim D}[h(x) \neq f(x)]$$

Learners differ in the way by which they select the classifier function. One plausible, and perhaps most intuitive, way of doing this selection is by simply choosing the classifier which minimizes risk, or agrees the most with f. This common learning paradigm is known as empirical risk minimization (ERM).

Some of the most interesting results of SLT have to do with predicting how much training data is needed to obtain a classifier that outputs the correct label to a desired level of accuracy. For example, if we assume that a perfect classifier f exists for a given classification problem with H being the hypothesis class under consideration, for any choice of $\delta \in (0, 1), \epsilon > 0$, and for an integer m satisfying $m \geqslant \frac{ln(|H|/\delta)}{\epsilon}$ (interpreted as the cardinality of the training set), running an ERM-learner over H guarantees that:

$$L_D(h_S) \leqslant \epsilon$$

with at least probability $1 - \delta$. Though this is a relatively rudimentary result of SLT, it is illuminating in that it reveals just how much training data is needed in order to obtain a classifier of a desired level of accuracy (represented by the ϵ term) to a desired degree of probability (represented by the δ term). With parameters for both accuracy and probability present, this result demonstrates what is canonically referred to in the literature as probably-approximately correct (PAC) learning, defined more precisely below:

All in all, in SLT, learning is interpreted as an agent selecting a maximally accurate classifier function, based on some amount of training data provided. As classification and regression problems are ubiquitous in the real world (e.g. in medicine, classifying tumors, or by banks, deciding whether or not to grant a loan), statistical learning has the benefit of being quite practically useful, even if learners do not converge to conclusions (i.e. the correct classifier function) with certainty. Developing and expanding SLT would therefore amplify these benefits.

3 Dynamic Epistemic Logic

Dynamic Epistemic Logic (DEL) was developed in order to deal with change in epistemic models. Specifically, in a multi-agent scenario, with each agent's knowledge and beliefs expressed by epistemic logic formulas, DEL is able to describe how the knowledge and beliefs of the agents change after an event.

To describe DEL, it may first be more fruitful to delve briefly into epistemic logic, of which DEL is eponymously the dynamic version. Epistemic logic was constructed to formally model epistemic principles and to explore their implications [Rendsvig et al., 2023]. Such epistemic principles are typically characterized by statements about knowledge and belief. For example, the well known axioms:

$$KB1 : K_a\varphi \to B_a\varphi$$
$$KB2 : B_a\varphi \to K_aB_a\varphi$$

which stand for "agent a knowing φ implies agent a believes φ" and "agent a believing φ implies agent a knows that they believe φ" respectively.

There are at least a couple of reasons we should want to use this sort of logical framework to model knowledge and belief. One is for precision. This language can capture precisely what an agent knows or believes and what it does not, and allows one to deduce additional statements about this with certainty. For instance, knowing with certainty that agent A knows φ, knowing that knowing φ implies knowing ψ, by KB1, we can deduce with certainty that A also believes ψ. Such certainty about knowledge could conceivably be useful in game-theoretic scenarios, where having information about what different agents know and believe could help predict their actions, thus enabling a given agent to pick the highest utility action for themselves. The semantics of epistemic logic (the dynamic version of which we will see shortly), being based on possible worlds, also coheres nicely with the developments made in late 20th century philosophy, of which possible worlds form a large theme.

Specifically, the epistemic logic formulas are evaluated on Kripke models.

Definition 3.1 (Kripke model) A standard Kripke model is a triple

$$M = (S, \{R_i\}_{i \in I}, ||.||, s_*),$$

with S representing a set of possible worlds, $\{R_i\} \subseteq S \times S$ a family of binary accessibility relations (indexed by labels $i \in I$), $||.|| : \Phi \to P(S)$ a valuation assigning to each $p \in \Phi$ a set $||p||_M$ of states, and s_* the actual/designated world.

In the context of epistemic logic and in a single-agent setting, there are just two accessibility relations in the family: ones representing knowledge and belief, which may be notated as $\sim$ and $\to$, respectively. For two worlds $a, b \in S$, $a \sim b$ may be interpreted as an agent not being able to distinguish a from b if a were the true state. Multi-agent Kripke models $M = (S, \{\xrightarrow{a}\}_{a \in A}, ||.||, s_*)$ are similar to the single agent version, apart from the accessibility relations, which are modified to reflect the knowledge of merely one given agent $a \in A$ in the multi-agent scenario.

As foreshadowed by the name, DEL is the dynamic version of epistemic logic, and is able to model updates to the situation, in terms of agents gaining new information. More specifically, where some new information φ is learned with certainty, the update $!\varphi$ is performed on the model, which corresponds to eliminating all the states in which φ does not hold [Baltag, 2023a]. That is, the new set of worlds after the update is:

$$||\varphi_{\mathbf{S}}|| = \{w \in S : w \models_{\mathbf{S}} \varphi\}$$

DEL is perhaps best illustrated further via use of examples. Public announcement logic (PAL) is one such example of the more general DEL we have thus far sketched out, and is used to reason about changes in knowledge and belief of agents after (assumed) truthful announcements have been made. To set up an example, PAL is defined by adding modalities to basic multi-modal logic:

Definition 3.2 (Syntax of PAL)

$$\varphi := p \mid \neg\varphi \mid \varphi \vee \psi \mid K_a\varphi \mid [!\varphi]\psi$$

with $p \in Prop$ and $a \in \mathcal{A}$, where $\mathcal{A}$ is the set of agents, and $K_a\varphi$ means that "agent a knows φ, and $[!\varphi]\psi$ stands for the dynamic update (to be explained more precisely). The respective semantics are found below:

Definition 3.3 (Semantics of PAL)

- $M, w \models p$ iff $w \in V(p)$
- $M, w \models \neg\varphi$ iff not $M, w \models \varphi$
- $M, w \models \varphi \vee \psi$ iff $M, w \models \varphi$ or $M, w \models \psi$
- $M, w \models K_i\varphi$ iff for every $w' \in St$ such that $w \sim_i w'$ we have $M, w' \models \varphi$
- $M, w \models [!\varphi]\psi$ iff if $\mathcal{M}, w \models \varphi$ then $\mathcal{M} \mid \varphi, w \models \psi$

To illustrate PAL in action, we consider the canonical muddy children puzzle. The description is as follows [Baltag and Renne, 2016]:

The Muddy Children Puzzle: Three children are playing in the mud. Father calls the children to the house, arranging them in a semicircle so that each child can clearly see every other child. "At least one of you has mud on your forehead", says Father. The children look around, each examining every other child's forehead. Of course, no child can examine his or her own. Father continues, "If you know whether your forehead is dirty, then step forward now". No child steps forward. Father repeats himself a second time, "If you know whether your forehead is dirty, then step forward now". Some but not all of the children step forward. Father repeats himself a third time, "If you know whether your forehead is dirty, then step forward now". All of the remaining children step forward. How many children have muddy foreheads?

The situation can be modeled using PAL, specifically in the form of a directed graph, where each node represents a possible world (in that world "cdd", with "c" representing "clean" means the foreheads of children 2 and 3 are muddy, while "ccc" means that the heads of all children are clean, and so on), and each bidirectional edge with label $l = i \in \{1, 2, 3\}$ between nodes signifies that the two worlds w_k, w_j between which the edge is exists are indistinguishable to child l. Further, d_i in this setup is the (atomic) PAL formula representing the proposition "child i is dirty" and each world satisfies the respective conjunction of such atoms (e.g. $\neg d_1 \wedge d_2 \wedge d_3$ is true at world cdd). If we assume that both children 1 and 2 have mud on their foreheads – that is, that "ddc" is the actual world, satisfying $d_1 \wedge d_2 \wedge \neg d_3$, before the first announcement, the initial model is rendered in Figure 2.

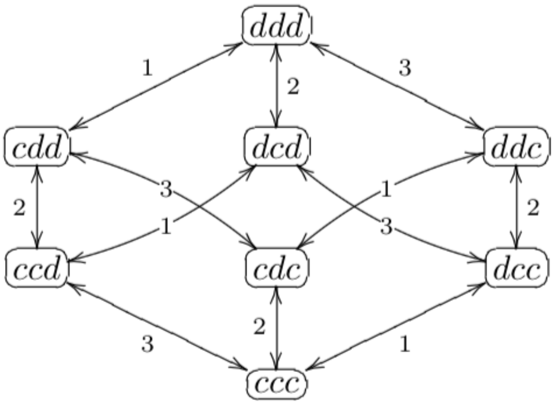

Figure 2: The initial model of the Muddy Children Puzzle, before any announcements have been made.

The first announcement constitutes an update to the model, in turn eliminating some of the possible worlds and edges from consideration for being the actual world. Specifically, it can be seen as an update $!(d_1 \vee d_2 \vee d_3)$ that results in the model in Figure 3.

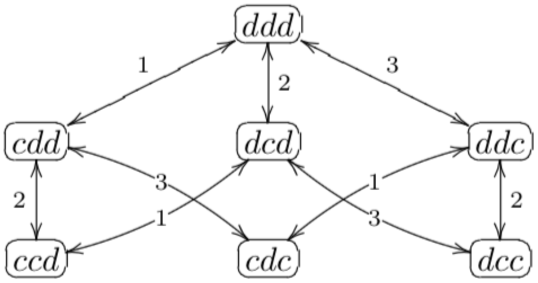

Figure 3: A model of the Muddy Children Puzzle after the first announcement.

For brevity, we do not consider how the rest of the scenario plays out, but eventually in continuing updates in this fashion, only the true world will remain in the model. In other words, PAL (and therefore, DEL) can model how a situation involving information change progresses, specifying what each agent knows at which juncture, and can be used to converge on the true world and the propositions it contains (assuming the updates or public announcements are themselves truthful).

To step back a bit more, the novel contribution of DEL could be then thought of as being a tool that enables one to converge to reason with certainty, as it pertains to matters about the external world (e.g. which children have muddy foreheads) and the knowledge and beliefs of others. This ability to model mental states of agents and to represent changes in these over time is specifically what distinguishes DEL from other logics, attributing to it the labels "epistemic" and "dynamic" respectively. This, in turn, can be useful, because insofar as agents strive to maximize utility in real-world settings, and that the real world contains many different agents whose goals may be at odds with each other, being able to accurately keep track of knowledge based on public announcements can guide the agent on how to act.

Some key questions the study of DEL seeks to answer are:

1. What are the best ways to logically represent common and distributed knowledge?
2. How should actions or announcements in the world affect one's knowledge and beliefs? What are implications of different types of actions (e.g. public vs. private, truthful vs. deceptive)?
3. How can DEL be integrated with other logical systems, such as temporal and modal logic?

In DEL, therefore, learning is the process of an agent gaining new information about the world in the form of logical propositions, unlike the learning of a classifier function in SLT. Specifying how to translate information from the world into logical form seems intuitively more challenging than encoding it as vector components as in SLT, but DEL has the advantage of representing knowledge of an agent as being certain, whereas in SLT agents only "know" propositions with a probability. Yet, the precise range of problems that DEL is applied to are seemingly narrower than SLT.

4 From statistical learning theory to formal learning theory

The goal of this and the subsequent sections is to attempt to bridge the mathematical setups of SLT and DEL. We will do so by first bridging SLT and formal learning theory (FLT). We propose that if we can bridge the setups of SLT and FLT, we could use the already established link between FLT and DEL [Gierasimczuk, 2009] to, in turn, bridge SLT and DEL as is illustrated in Figure 1. FLT was first presented in [Gold, 1967], a field investigating inductive learnability of concepts in the limit. [Baltag, 2023b] motivated FLT laconically:

> "Why formal learning theory? Flexible and open-ended approach for inductive learning from successful observations. While most other approaches adopt a normative stance, FLT gives the learner a high degree of freedom, allowing the choice of any learner that produces conjectures based on the data (no matter how 'crazy' or unjustified are these conjectures, or how erratic is the process of belief change)."

While [Gierasimczuk et al., 2014] described the process of data aggregation in FLT comprehensively:

> "To illustrate, consider a game between a learner and nature (or teacher) where the learner needs to identify the current state of the world. We assume that the

> incoming information is readable and that all the data that are consistent with the actual world are eventually presented to the learner. The source of data is also taken to be truthful (since nature never lies). This game is described as follows. Initially, there is a class of concepts (or class of realities). Intuitively, this class represents the uncertainty range of the learner. Nature chooses at the beginning of the game one of these concepts to be the target concept and starts providing to the learner pieces of data concerning the target concept. The learner's aim is to guess correctly which concept from the class is the one chosen by nature. If the learner succeeds, we say that the learner identifies (or learns) the target concept. If the learner identifies every concept of the class, we say that the learner identifies the class of concepts."

Our point of departure in this section will be the works [Gierasimczuk, 2009] and [Gierasimczuk et al., 2014]. These papers investigated the connection between inductive learning and DEL which is convenient for our purposes because SLT is a paradigm for studying inductive learning while DEL, traditionally, is not.

There are a few things lacking in the mentioned works, however – to be able to fully transport the questions, objects, and results from SLT to DEL. Specifically:

- DEL primarily deals with aggregating incoming information about one scenario or data subject, its internal model is updated one feature/epistemic fact at a time. SLT, on the other hand, processes various features of the data subject "in one go", usually all of the features of the data subject's are represented in a single vector.
- The outcome of a DEL learning process is a Kripke model with the agents' beliefs about the true situation and is traditionally not used to predict the outcome of new data. On the contrary, the outcome of an SLT learning process is a function that fits the past data and can be used for predicting the outcome of new data.
- DEL can deal with multiple agents, in SLT there are no agents. At best, in SLT we can identify a single agent - the learner.
- DEL deals with modelling the knowability and belief of agents, SLT cannot model such scenarios, it primarily deals with predicting an integer (classification task) or a real number (regression task).
- In DEL, there is no quantitative measure of success, only subjectively qualitative ("what does each agent know and believe?", "how can we use this Kripke model, having incorporated the epistemic facts?"). In SLT, we have a well-formalized notion of risk and accuracy which serve as the quantitative measures of success.

The aforementioned differences are substantial, but not impossible to overcome. This is another reason why FLT may prove useful. Forging a link between SLT and DEL directly seems like a difficult task, while doing that for SLT and FLT first is more straightforward. Also, in doing so, we can figure out which of these issues are actually worrisome when trying to bring SLT and DEL frameworks together.

We now introduce the main definitions of FLT which are based on [Gierasimczuk et al., 2014] and [Baltag, 2023b].

Definition 4.1 (Epistemic space) Let (S, Φ) be an epistemic space. Here, S is the set of epistemic possibilities (worlds) and $\Phi \subseteq \mathcal{P}(S)$ a family of propositions. The propositions represent facts or observables being true or false in any of the possible worlds under consideration.

Definition 4.2 (Data stream) A **data stream** is an infinite sequence of propositions from Φ, that is, $\overrightarrow{O} = (O_1, O_2, \ldots)$ which is assumed to be consistent: $\bigcap_{i=1}^{\infty} O_i \neq \emptyset$.

A data sequence is any finite initial segment $\overrightarrow{O}[n] = (O_1, \ldots, O_n)$ of a stream.

A data stream $\overrightarrow{O}$ is **sound** w.r.t. a world s iff every data observed in $\overrightarrow{O}$ is true at s, that is, for all n, $s \in O_n$.

A data stream $\overrightarrow{O}$ is **complete** w.r.t. a world s iff every observable property of s appears in $\overrightarrow{O}$ is true at s: $\forall n, s \in O_n$.

A data stream $\overrightarrow{O}$ is **a data stream for** s if it is both sound and complete w.r.t. s.

Definition 4.3 (Learning method (FLT)) A Learning method $\mathcal{L}$ is a function that on input of an epistemic space and a finite sequence of observations $\overrightarrow{O}[n] = (O_1, \ldots, O_n)$ outputs a hypothesis. The **hypothesis** is then a set of possible worlds, i.e, a proposition, so $\mathcal{L}((S, \Phi), \overrightarrow{O}[n]) \subseteq S$.

To establish a formal analogy between learning problems in SLT and FLT, a connection between their core mathematical constructs must be demonstrated. We will first present notions in SLT and then discuss their parallels in FLT. After a brief philosophical discussion, we will present a formal definition.

What is the space of epistemic possibilities? At first, the SLT learner considers the whole hypothesis space $\mathcal{H}$ which is a set of functions. This function space is reliant on the learning algorithm that we assume, but for this paper we side-step this detail. Since functions are fully described and differentiated in terms of their input-output behaviour, they resemble epistemic possibilities, which share that quality - they are characterized by the propositions which are true at them. So, we can set $S = \mathcal{H}$ for the conversion.

What are the observables, the data not necessarily available to the learner but possible in theory? In SLT that is the true probability distribution D defined on some domain $\mathcal{X}$. It is also assumed that some true function f exists which labels the data points in some way. In FLT, that is the family of propositions $\Phi \subseteq \mathcal{P}(S)$, it represents all observable facts *in theory*. We can thus simply set $\Phi = \mathcal{P}(\mathcal{H})$, each observable $O \in \Phi$ would represent a collection of functions which, intuitively, means that O is the set of functions that fit the data perfectly.

What are the learner's hypotheses? The SLT learner, upon seeing some training data S, returns a single hypothesis function $h \in \mathcal{H}$. Note that there may be more hypotheses that fit the training data perfectly, but the learner outputs one of them. In a similar manner, the FLT learner, after observing a data sequence $\overrightarrow{O}[n]$, it returns a hypothesis $h \subseteq S$ which is a set of epistemic possibilities, representing the set of functions which fit the data seen. While the SLT learner returns a single output and the FLT learner returns a set, we think it is intuitively correct and could be mended with small modifications to either of the frameworks.

What data is available to the learner? The SLT learner receives a finite training set which is of the form $S = \{s_1, s_2, \ldots, s_n\} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$. Each s_i is a tuple where the first element $\mathbf{x}_i$ is a multi-dimensional vector describing the properties of the object to classify and the second element $y_i \in \{0, 1\}$ is labeled by some "correct" function f. The whole learning task in SLT is to approximate f as best as possible. The training set S is sampled from an arbitrary distribution D. The amount of data points that we can sample is usually dictated by the situation at hand as the amount of data available is an important topic of investigation.

In FLT, the data sequence $\overrightarrow{O}[n]$ is a finite consistent set of propositions. We can craft a corresponding $\overrightarrow{O}[n]$ using the training sample S by simply taking each O_i to be $\arg\min_{h \in \mathcal{H}} L_{s_i}(h)$, that is, the set of functions which minimize the risk with respect to one data point s_i. Note that $\overrightarrow{O}[n]$ is indeed consistent (the intersection of all functions is non-empty) as some of them must agree on how to fit certain data points.

To view a summary of these definitions at a glance, see Table 1.

Definition 4.4 (SLT model to FLT model) Take an SLT model $(\mathcal{H}, S, L)$ which is comprised of a hypothesis space $\mathcal{H}$, training sample S (which is sampled from some true distribution D), and risk function L. The epistemic space is then a tuple $(\mathcal{H}, \mathcal{P}(\mathcal{H}))$ and an arbitrary data stream is $\overrightarrow{O} = (O_1, O_2, \ldots)$ where each $O_i = \arg\min_{h \in \mathcal{H}} L_{\{s_i\}}(h)$, that is, each element O_i is a set of hypotheses which minimizes the risk on that particular data sample s_i.

In practice, the hypothesis space is completely dependent on the learning algorithm we pick (logistic regression, decision tree, neural networks, etc.). However, this conversion is general enough to account for that as well as the type of distribution we sample from or the concrete risk function we pick to calculate the errors.

Let us illustrate the conversion with a practical example. We will describe the initial learning scenario in the SLT framework and then perform the conversion. We will then use the respective learning algorithms on both results separately and show that the results are the same. This example will (informally) illustrate that the conversion is correct.

Example 4.1 Let $\mathcal{H} = \{a, b, c, d, e, f, g\}$ be the set of hypothesis functions, the hypothesis space. Let the training set be

$$S = \{(\mathbf{x}_1, -), (\mathbf{x}_2, -), (\mathbf{x}_3, +), (\mathbf{x}_4, -), (\mathbf{x}_5, +)\}$$

Statistical Learning Theory		**Formal Learning Theory**	
Name	**Notation**	**Notation**	**Name**
Hypothesis space	$\mathcal{H}$	S	Epistemic possibilities
All subsets of hypotheses	$\mathcal{P}(\mathcal{H})$	Φ	Observables
Min-risk functions on single training examples	$\forall s_i, \arg\min_{h\in\mathcal{H}} L_{\{s_i\}}(h)$	$\overrightarrow{O}[n]$	Data sequence
Learning algorithm	A	$\mathcal{L}$	Learner
Empirical risk; True risk	$L_S; L_D$	-	No *degrees* of success

Table 1: This table demonstrates the correspondence between various mathematical objects employed in Statistical Learning Theory and their close parallels within Formal Learning Theory.

where $-$ and $+$ are the negative and positive labels of the samples respectively, labeled by the true labeling function f. Let L be the usual count-based ratio risk. Assume the learning algorithm A to be empirical risk minimizer as described in section 2.

Suppose that the hypotheses fit the training set as follows. Given the first data point, four functions out of seven (the whole hypothesis space) fit it, namely, hypotheses b, d, e, g. Hypotheses b, d, f, g fit the second data point. Hypotheses a, c, d, e fit the third sample. Hypotheses a, b, c, d, e, g fit the fourth point. Hypotheses b, c, d fit the fifth point. This process of which hypotheses fit the data are illustrated in Figure 4 through Figure 8.

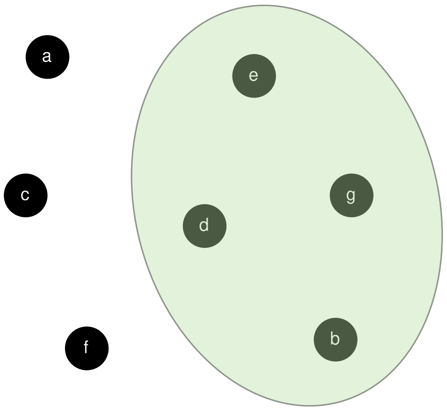

Figure 4: The first data point from the training sample is fit by hypotheses b, d, e, g.

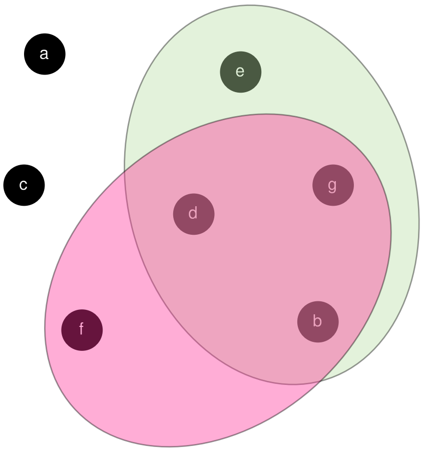

Figure 5: The second data point from the training sample is fit by hypotheses b, d, f, g.

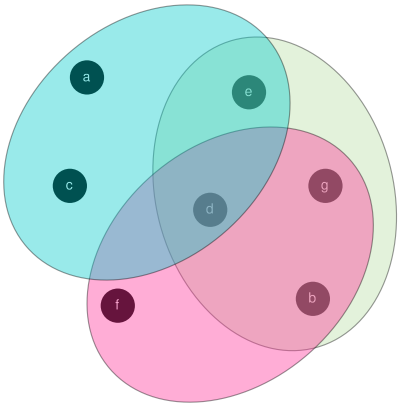

Figure 6: The third data point from the training sample is fit by hypotheses a, c, d, e.

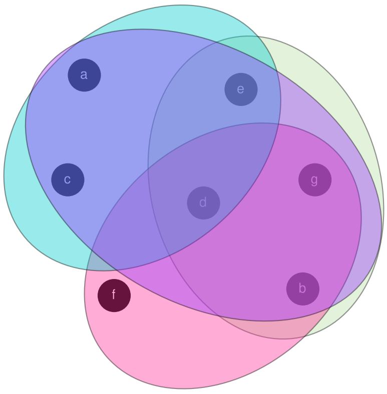

Figure 7: The fourth data point from the training sample is fit by hypotheses a, b, c, d, e, g.

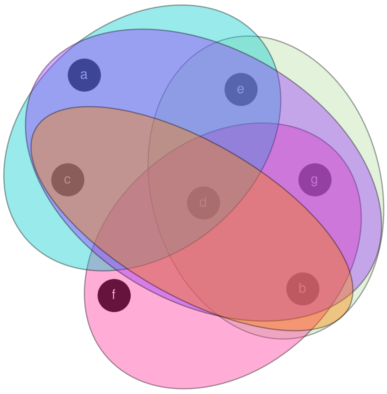

Figure 8: The fifth data point from the training sample is fit by hypotheses b, c, d.

Because of these facts, the empirical risks of the each of these functions are as follows:

- $L_S(a) = \frac{3}{5} = 0.6$. This is because 3 data points out of the training sample of 5 were not fit by a.
- $L_S(b) = \frac{1}{5} = 0.2$
- $L_S(c) = \frac{2}{5} = 0.4$
- $L_S(d) = \frac{5}{5} = 0$
- $L_S(e) = \frac{2}{5} = 0.4$
- $L_S(f) = \frac{4}{5} = 0.8$
- $L_S(g) = \frac{2}{5} = 0.4$

So, since the learning method is ERM, we will clearly get that $A(S) = \{d\}$ since its empirical risk is the lowest, i.e., $d \in \arg\min_{h\in\mathcal{H}} L_S(h)$.

Now, construct the corresponding epistemic space $(\mathcal{H}, \mathcal{P}(\mathcal{H}))$ as well as the sequence based on the training set: $\overrightarrow{O}[5] = (O_1, O_2, O_3, O_4, O_5)$. As mentioned in Definition 4.4, each $O_i = \arg\min_{h\in\mathcal{H}} L_{s_i}(h)$ for each $s_i \in S$. For example, $O_1 = \arg\min_{h\in\mathcal{H}} L_{s_1}(h) = \{b, d, e, g\}$.

It is easy to see (e.g. from Figure 8) that we can set the FLT learner $\mathcal{L}$ to be simply intersection of all observations, namely:

$$\mathcal{L}((\mathcal{H}, \mathcal{P}(\mathcal{H})), \overrightarrow{O}[5]) = \bigcap_{i\in\{1,\dots,5\}} O_i = \{d\}$$

♠

5 From statistical learning theory to dynamic epistemic logic

In the previous section, we introduced FLT and conjectured a bit on how a statistical learning theory model can be converted to an FLT model. We will now use these results to finish the bridge between SLT and DEL. The main result of this section (and paper) will be an extended theorem from [Gierasimczuk et al., 2014]. First, however, we must introduce the necessary concepts of finite identification and a conversion from FLT to DEL models.

Definition 5.1 (Finite identification) Let (S, Φ) be an epistemic space. A learning method $\mathcal{L}$ finitely identifies $s \in S$ if, for every stream $\overrightarrow{O}$ for s, there exists $n \in \mathbb{N}$ such that $\mathcal{L}((S, \Phi), \overrightarrow{O}[n]) = \{s\}$ for all $k \geqslant n$ and $\mathcal{L}((S, \Phi), \overrightarrow{O}[n]) = ?$ for all $k < n$, where ? is the learner withholding judgement, outputting "I don't know".

The epistemic space (S, Φ) is said to be **finitely identifiable** by $\mathcal{L}$ if all its worlds are finitely identifiable by $\mathcal{L}$.

The epistemic space (S, Φ) is **finitely identifiable** just in case there is a learning method that can finitely identify it.

As we will see below, this property will be central to the bridging of some concepts between our learning frameworks in question.

It remains to demonstrate how inductive learning is modeled in DEL, or how one converts an FLT learning model to a DEL model, as described in [Gierasimczuk et al., 2014]. Conversions like these are necessary to model concepts from one framework in an another framework. The intuition of this conversion is described below.

Let (S, Φ) be an epistemic space. Take the initial class of sets S to be possible worlds in an epistemic model. This will reflect the learner's initial uncertainty over the range of sets.

The observations that the learner is exposed to will be the events that modify the initial model.

The agent's uncertainty is dynamically refined through the incorporation of new data. Data is presented as propositions received from a source assumed to be completely truthful. In this context, the agent employs a logic of propositional and epistemic update to eliminate possible worlds that are inconsistent with the received information. This approach aligns with learning theory, where the veracity of incoming data is often a foundational assumption, justifying the use of propositional and epistemic update as a framework for inquiry.

Formally, the transformation from formal learning theory models to dynamic epistemic models is defined in the following manner [Gierasimczuk et al., 2014].

Definition 5.2 (FLT model to DEL model) Take an epistemic space (S, Φ). For every proposition in $p_n \in \Phi$ we take a symbol $\mathtt{p}_n \in \mathit{Prop}$. Moreover, take a set Nom which contains a nominal symbol $\mathtt{i}$ for every $i \in \mathbb{N}$. The initial learning model $\mathcal{M}_{(S,\Phi)}$ is a triple $\langle W, \sim, V \rangle$, where $W := S, \sim := W \times W$, $V : \mathit{Prop} \cup \mathit{Nom} \to \mathcal{P}(W)$, such that $s_i \in V(\mathtt{p}_n)$ iff $s_i \in p_n$ in (S, Φ), and for any $\mathtt{i} \in \mathit{Nom}$ we set $V(\mathtt{i}) = \{s_i\}$.

The following theorem is the central result linking FLT and DEL. It was first proved in [Gierasimczuk, 2009], but this formulation is taken from [Gierasimczuk et al., 2014].

Theorem 5.1 The following are equivalent:

1) An epistemic space (S, Φ) is finitely identifiable.

2) For every $s_i \in S$ and every data stream $\overrightarrow{O}$ for s_i there is an $n \in \mathbb{N}$ such that for all $m \geqslant n$, $\mathcal{M}_{(\mathcal{S},\Phi)}, s_i \models \left[!\left(\bigwedge \mathit{set}\left(\overrightarrow{O}[m]\right)\right)\right] K\mathtt{i}$.

Where $\mathit{set}(\cdot)$ operator converts any object into a set (in this case, converts a sequence to a set).

Part 2) of this theorem is explained by [Gierasimczuk et al., 2014]:

> "In terms of propositional knowledge and belief this corresponds to the following: whatever is true in the actual world I know (believe) that it is true and vice versa. In other words, we may say: $K\mathtt{i}$ iff for any $o \in \mathit{Prop}$ such that $s_i \in p$ we have that $p \iff Kp$."

We finally have all the ingredients necessary to provide an original contribution. As we've been building up throughout this paper, we would like to link some existing result from statistical learning theory to part 1) of the aforementioned theorem thus showing that some part of SLT can be modeled in FLT (and DEL).

In the context of FLT, finite identification refers to the capability of a learning algorithm to definitively identify an object within a finite number of data presentations [Gold, 1967].

This implies the algorithm's ability, for any stream of data for the object in question, to halt data acquisition after a specific point, having accumulated sufficient evidence to correctly identify the object. This is a strong condition concerning arbitrary streams of data for an object.

In SLT, the notion of uniform convergence of a hypothesis space resembles this. The uniform convergence condition is defined formally in [Shalev-Shwartz and Ben-David, 2014, p. 55] as follows.

Definition 5.3 (Uniform convergence) A hypothesis class $\mathcal{H}$ has the uniform convergence property if there exists a function $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ such that:

For every $\epsilon, \delta \in (0,1)$, for every distribution D over $\mathcal{X}$, if S is a sample of $m \geqslant m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d examples drawn from D, then with probability $1 - \delta$,

$$\sup_{h \in \mathcal{H}} |L_S(h) - L_D(h)| \leqslant \epsilon$$

Uniform convergence, similarly like finite identification, is also a strong condition concerning the theoretically available data for the hypotheses. Informally, uniform convergence says that each hypothesis in the hypothesis space has both low empirical risk as well as low true risk if enough samples are given to train on. The precise number of samples for what is "enough" to ensure low empirical and true risks may be the *finite number* which would enable finite identification of corresponding epistemic spaces. We will prove this as the next theorem.

And so, the main result of this paper which shows that some results of SLT can be transported to DEL is extending the Theorem 10.6 from [Gierasimczuk et al., 2014] with an additional equivalence result:

Theorem 5.2 The following are equivalent

1) A hypothesis space $\mathcal{H}$ has the uniform convergence property with sample complexity $m_{\mathcal{H}}^{UC}(0,0)$.

2) An epistemic space $(\mathcal{H}, \mathcal{P}(\mathcal{H}))$ is finitely identifiable.

3) For every $h_i \in \mathcal{H}$ and every data stream $\overrightarrow{O}$ for h_i there is an $n \in \mathbb{N}$ such that for all $m \geqslant n$, $\mathcal{M}_{(\mathcal{H}, \mathcal{P}(\mathcal{H}))}, h_i \models \left[! \left(\bigwedge set \left(\overrightarrow{O}[m] \right) \right) \right] K\mathtt{i}$.

Proof. As mentioned, the statements 2) → 3) and 3) → 2) were proved in [Gierasimczuk, 2009]. Now, we focus on proving 1) → 2) and 2) → 1). If this is proved, then we can prove 1) → 3) by simply transforming the SLT model to FLT and, in turn, to a DEL model and vice-verse for 3) → 1).

First, let us prove 1) → 2). Assume we have a hypothesis space $\mathcal{H}$ which has the uniform convergence property with sample complexity $m_{\mathcal{H}}^{UC}(0,0)$. That is, for any distribution D and $\epsilon = 0, \delta = 0$, we have

$$\sup_{h \in \mathcal{H}} |L_S(h) - L_D(h)| \leqslant 0$$

with probability $\geqslant 1-\delta = 1-0 = 1$. So, if *any* learning algorithm observes $m \geqslant m_{\mathcal{H}}^{UC}(0,0)$ data points, it is guaranteed to output a hypothesis which makes no prediction mistakes, even when predicting unobserved samples.

Now, convert this SLT scenario to an epistemic space $(\mathcal{H}, \mathcal{P}(\mathcal{H}))$ as in Definition 4.4. It remains to show that the epistemic space $(\mathcal{H}, \mathcal{P}(\mathcal{H}))$ is finitely identifiable, i.e., there is a learning method $\mathcal{L}$ such that finitely identifies all $h \in \mathcal{H}$. Recall that a learning method $\mathcal{L}$ finitely identifies h if, for every stream $\overrightarrow{O}$ for h, there exists a finite number m of observations such that $\mathcal{L}((\mathcal{H}, \mathcal{P}(\mathcal{H})), \overrightarrow{O}[m]) = \{h\}$.

Since the hypothesis space has the uniform convergence property, which is a property of the hypothesis space itself and not of the learning method, it allows us to take a very simple learner. Let $\mathcal{L}$ be a (FLT) learning method as follows:

$$\mathcal{L}\left((\mathcal{H}, \mathcal{P}(\mathcal{H})), \overrightarrow{O}[k]\right) = \begin{cases} ? & \text{if } k < m \\ \{h_m\} & \text{if } k \geqslant m \end{cases}$$

where $m = m_{\mathcal{H}}^{UC}(0,0)$ and h_m is the hypothesis which is fit on m examples. The intuition behind arriving at h_m is simply due to the fact that the sequence is of size at least m which "shrinks" the uncertainty substantially, the remaining hypothesis space is reduced to functions of minimal (in this case, 0) risk. So, $\mathcal{L}$ would output the correct hypothesis fitting the data $\overrightarrow{O}[k]$ perfectly if it observes at least m data points, otherwise, the learning method would withhold judgement. Since this is applicable for arbitrary streams $\overrightarrow{O}$ for any h, we can say that $\mathcal{L}$ finitely identifies $(\mathcal{H}, \mathcal{P}(\mathcal{H}))$.

We have shown that, due to uniform convergence, the hypothesis space not only has *some* finite $n \in \mathbb{N}$ for any $\overrightarrow{O}$ for any $h \in \mathcal{H}$, but a concrete integer $m \geqslant m_{\mathcal{H}}^{UC}(0,0)$ which allows us to construct a learning method to finitely identify the epistemic space $(\mathcal{H}, \mathcal{P}(\mathcal{H}))$.

□

Let us now turn to proving 2) → 1). Assume the epistemic space $(\mathcal{H}, \mathcal{P}(\mathcal{H}))$ is finitely identifiable. We need to show that there is some integer $m = m_{\mathcal{H}}^{UC}(0,0)$ such that, for all $h \in \mathcal{H}$, $\sup_{h\in\mathcal{H}}|L_S(h) - L_D(h)| = 0$. Since the epistemic space is finitely identifiable, there exists some learner $\mathcal{L}$ that finitely identifies it. Specifically, each of $h \in \mathcal{H}$ is finitely identifiable, and each stream for h has a finite $n \in \mathbb{N}$ such that the learner $\mathcal{L}$ correctly identifies h.

Let $M = \{n_1, n_2, \ldots\}$ where n_i is an integer representing a finite time after which h_i can be finitely identified by $\mathcal{L}$. Then, simply take $m = \max(M)$. Since the learning in FLT is deterministic, we have $\delta = 0$. Since we have assumed that $\overrightarrow{O}$ is consistent, we know that there will be no mistakes made by $\mathcal{L}$. In addition, the set M comprises the integers after which $\mathcal{L}$ definitely takes a decision, does not ever output ?, so that is not a bother for us when transferring the result to SLT. From these last two arguments we can conclude that we can set $\epsilon = 0$. Any $n_i \in M$ is an integer for *any* stream for h_i, so, on the SLT part, these results will work for any distribution D over the domain. And so, if S is a training set of size $m \geqslant \max(M)$ drawn from D, then with probability $1 - \delta = 1$ it follows that for any $h \in \mathcal{H}$,

$$\begin{aligned} L_S(h) = L_D(h) = 0 &\iff \\ L_S(h) - L_D(h) \leqslant 0 &\iff \\ \sup_{h\in\mathcal{H}}|L_S(h) - L_D(h)| \leqslant 0 = \epsilon \end{aligned}$$

Thus, $\mathcal{H}$ has the uniform convergence property with sample complexity $m = m_{\mathcal{H}}^{UC}(0,0)$. □

■

Essentially, the theorem above says that we can model uniform convergence (for a fixed sample complexity function) in FLT and DEL. If any uncertainty and error is removed from SLT learning scenarios, it starts to resemble the FLT learning scenarios. The abstract learnability results of SLT (e.g. every part of the fundamental theorem of statistical learning theory), just like the results of FLT, deal with a type of "learning in the limit". SLT deals with more realistic scenarios concerning finite data and puts much more focus on the specific amount of data necessary to draw conclusions, while FLT is a framework with less restrictions and often discusses learnability scenarios with infinite data. DEL, on the other hand, primarily deals with one-step revisions of its epistemic model, something that both vanilla SLT and FLT lack. However, in Theorem 5.2 we see that DEL is expressive enough to model the "in the limit" learning scenarios. Once we start to consider specific learning algorithms (e.g. decision trees, neural networks) which come with their own specific rules how they incorporate each training sample, the framework of SLT starts to resemble the DEL learning scenarios.

There is more work to be done to bring all these three frameworks together in order to crystallize their strengths and differences. We hope that in the process the scientific community will arrive at results which lead to new understandings of learnability.

6 Conclusion and future work

In this paper we set out to bring the frameworks of statistical learning theory and dynamic epistemic logic closer. In the process, we have also described how statistical learning theory learning models can be framed in terms of formal learning theory. We used the existing result from [Gierasimczuk, 2009] which presented a link between FLT and DEL. We presented a similar novel result to describe a bridge between SLT and FLT to, in turn, model some results, specifically, uniform convergence, in DEL.

Our work can be used as a cornerstone for many further investigations of the aforementioned learning and knowledge frameworks. Some concrete ideas for future work are as follows:

- How do hypotheses spaces that have uniform convergence with $m_{\mathcal{H}}^{UC}(\epsilon, \delta)$ for $\epsilon, \delta > 0$ transfer to FLT? To DEL models? Are the resulting DEL models similar to dynamic updates with probabilities [Van Benthem et al., 2009]?
- Is the realizability assumption important when transferring the results from SLT to FLT?

- What does the no-free-lunch theorem correspond to in FLT and DEL?
- What does VC dimension correspond to in FLT and DEL? What is the epistemic interpretation of VC dimension?
- There are strong connections between FLT and topology. How can we employ tools from topology to study concepts of SLT?
- Explore the implications of unique characterisability [Balder ten Cate, 2024] of worlds in FLT learning models.

References

[Balder ten Cate, 2024] Balder ten Cate (2024). Logic and Data Examples: Logical Foundations of Example-Driven Specification.

[Baltag, 2023a] Baltag, A. (2023a). Dynamic epistemic logic lecture slides 2.1.

[Baltag, 2023b] Baltag, A. (2023b). Topology, logic and learning lecture lecture slides 4.1.

[Gierasimczuk, 2009] Gierasimczuk, N. (2009). Bridging learning theory and dynamic epistemic logic. 169(2):371–384.

[Gierasimczuk et al., 2014] Gierasimczuk, N., Hendricks, V. F., and de Jongh, D. (2014). *Logic and Learning*, pages 267–288. Springer International Publishing, Cham.

[Gold, 1967] Gold, E. M. (1967). Language identification in the limit. 10(5):447–474.

[Shalev-Shwartz and Ben-David, 2014] Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

[Van Benthem et al., 2009] Van Benthem, J., Gerbrandy, J., and Kooi, B. (2009). Dynamic Update with Probabilities. 93(1):67–96.