Artificial Intelligence 174 (2010) 639–669

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Partial observability and learnability ☆

Loizos Michael

Open University of Cyprus, Cyprus

ARTICLE INFO

Article history:
Received 11 November 2007
Received in revised form 30 March 2010
Accepted 30 March 2010
Available online 1 April 2010

Keywords:
Partial observability
Appearance
Reality
Masking process
Sensing
Missing information
Autodidactic learning
Probably approximately correct
Information recovery
Reduction

ABSTRACT

When sensing its environment, an agent often receives information that only partially describes the current state of affairs. The agent then attempts to predict what it has not sensed, by using other pieces of information available through its sensors. Machine learning techniques can naturally aid this task, by providing the agent with the rules to be used for making these predictions. For this to happen, however, learning algorithms need to be developed that can deal with missing information in the learning examples in a principled manner, and without the need for external supervision. We investigate this problem herein. We show how the *Probably Approximately Correct* semantics can be extended to deal with missing information during both the learning and the evaluation phase. Learning examples are drawn from some underlying probability distribution, but parts of them are hidden before being passed to the learner. The goal is to learn rules that can accurately recover information hidden in these learning examples. We show that for this to be done, one should first dispense the requirement that rules should always make definite predictions; "don't know" is sometimes necessitated. On the other hand, such abstentions should not be done freely, but only when sufficient information is not present for definite predictions to be made. Under this premise, we show that to accurately recover missing information, it suffices to learn rules that are highly consistent, i.e., rules that simply do not contradict the agent's sensory inputs. It is established that high consistency implies a somewhat discounted accuracy, and that this discount is, in some defined sense, unavoidable, and depends on how adversarially information is hidden in the learning examples.

Within our proposed learning model we prove that any PAC learnable class of monotone or read-once formulas is also learnable from incomplete learning examples. By contrast, we prove that parities and monotone-term 1-decision lists, which are properly PAC learnable, are not properly learnable under the new learning model. In the process of establishing our positive and negative results, we re-derive some basic PAC learnability machinery, such as Occam's Razor, and reductions between learning tasks. We finally consider a special case of learning from partial learning examples, where some prior bias exists on the manner in which information is hidden, and show how this provides a unified view of many previous learning models that deal with missing information.

We suggest that the proposed learning model goes beyond a simple extension of supervised learning to the case of incomplete learning examples. The principled and general treatment of missing information during learning, we argue, allows an agent to employ learning entirely autonomously, without relying on the presence of an external teacher, as is the case in supervised learning. We call our learning model *autodidactic* to emphasize the explicit disassociation of this model from any form of external supervision.

☆ A preliminary version of this work appeared as: Loizos Michael, Learning from partial observations, in: Manuela M. Veloso (Ed.), Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI'07), January 2007, pp. 968–974. This work was completed while the author was at the School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA, and was supported by grant NSF-CCF-04-27129.

E-mail address: loizos@ouc.ac.cy.

0004-3702/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.artint.2010.03.004

1. Introduction

It can be argued that a central aspect of a fully autonomous agent is the ability to learn the rules that govern its environment, without any form of external supervision. An autonomous agent senses its environment and obtains information that is often incomplete, which serves, then, as input to the learning process. Such settings necessitate, thus, the use of learning algorithms that can deal with such incomplete learning examples.

In this work we propose a framework within which learning from incomplete learning examples can be formally studied. For concreteness, our framework can be viewed as an extension of the Probably Approximately Correct semantics [28]. Our goal is to show that it is possible to learn rules that accurately predict information missing in an agent's sensory readings, and that these rules can be obtained efficiently, and be accompanied by formal PAC-like guarantees, irrespectively of how information is hidden in the learning examples available during the learning and evaluation phases. We note, however, and point out throughout this work, that the problem of learning from incomplete learning examples goes beyond learning classification rules as in the original PAC model. We view the results of this work as a first step towards the more ambitious goal of devising learning algorithms that can identify more general rules.

Our exposition starts in Section 2, where the problem of learning from incomplete information is put into context, as the problem underlying the process of scientific discovery: identifying the structure of some underlying reality, given only partial appearances of that reality. We continue to show how the PAC semantics can be extended to this effect. As in the PAC model, learning examples are drawn independently at random from some underlying probability distribution. Unlike the PAC model, these examples are never directly accessible by an agent. Instead, some *arbitrary* stochastic process hides parts of these examples, giving rise to what we call *partial observations*. These observations are then given to the agent, both during the learning phase as a means to facilitate learning, and during the evaluation phase as the input on which learned rules are to be applied to make predictions, and against which these predictions are to be tested.

Due to lack of complete information during the evaluation phase, we allow learned rules to make "don't know" predictions, but only when the rules cannot be unambiguously evaluated on a given observation. Under this provision, we define a rule to be *consistent* with an observation if the rule's prediction does not directly contradict what is stated in the observation. In particular, if the observation does not offer any information on some target attribute, then any prediction is consistent. Learning is successful if highly consistent rules can be obtained efficiently in the relevant learning parameters.

We then consider a stronger notion of learnability, that of deriving rules that make predictions in a manner not only consistent with an observation, but *accurate* with the underlying example. Thus, even if the observation does not offer any information on some target attribute, the prediction may be accurate or not depending on what the hidden underlying value of the target attribute is. We show that this more stringent notion of learnability is information-theoretically unattainable when information is hidden adversarially in observations. We introduce a metric called *concealment* to capture the extent of this adversity, and show that consistency, accuracy, and concealment are tied together in a natural manner: consistency implies accuracy discounted by some factor determined by the concealment. This allows us to focus on the conceptually simpler notion of consistent learnability for the remaining of this work.

Section 3 discusses some of the choices we have made in our learning model, and contrasts them against existing work in Statistical Analysis and Learning Theory. Three main aspects are discussed: (i) when are "don't know" predictions allowed, and what does it mean to predict "don't know"; (ii) to what extent is autonomy possible when learning; and (iii) how much regularity is assumed in the way information is missing in learning examples. This discussion shows, in particular, that unlike most previous work, our learning framework does without the assumption of an external teacher. We call the learning framework *autodidactic* in recognition of this property.

The two subsequent sections provide positive and negative learnability results for autodidactic learnability. Section 4 establishes that certain machinery available in the PAC model applies also, in some form, in the context of autodidactic learnability. In particular, Occam's Razor [4] applies unchanged as in PAC learnability, while reductions between learning tasks [23] can be formalized in a way that accommodates the more stringent requirements that need to be met for autodidactic learnability. Using reductions we then establish that any PAC learnable concept class that contains only monotone or read-once formulas can also be learned autodidactically. Hence, in a broad set of domains, the lack of complete information does not render learnability any harder. By contrast, Section 5 establishes that incomplete information may in some cases diminish learnability. We show that, although they are properly PAC learnable, the concept classes of parities and monotone-term 1-decision lists are not properly learnable in the autodidactic model, unless $\mathtt{RP} = \mathtt{NP}$.

The case where information is not hidden completely arbitrarily in learning examples is examined in Section 6. We argue, and demonstrate, that depending on how structured such information hiding is, the semantics of missing information may be significantly altered, to the extent that missing information may, even, make learnability easier than in the case of complete information. Related learning models are then presented along with the assumptions they make on this structure.

We conclude in Section 7 with a list of open problems, and some pointers to future work.

2. Autodidactic learnability

The dichotomy between appearance and reality is inherent in the process of scientific discovery. Appearances are partial depictions of the reality that governs our world, and through such appearances scientists attempt to derive a model, or hypothesis, of the structure present in the underlying reality. The hypothesis is then applied on these appearances to make predictions about unobserved properties of the world. In physics, for instance, these predictions might concern spatially or temporally distant properties of the world, for which readings cannot be obtained through our sensors. Central in this process seem to be certain premises:

(i) Structure exists in the underlying reality of the environment, and not necessarily in the way that sensors hide information about this reality to give rise to appearances.
(ii) This underlying structure cannot be learned if it remains perpetually inaccessible through sensing.
(iii) Any attempt to discover this structure should rely solely on the partial information of the structure that is provided through the sensors, without any external supervision during the learning phase.
(iv) Developed hypotheses aim to model the structure of the underlying reality, and not necessarily the way that sensors hide information about this reality.
(v) A hypothesis about the underlying structure is applied to predict some of the missing information not present in appearances, given only whatever other partial information is available in appearances.

Machine learning research seems to have largely ignored these premises. In the words of McCarthy [18]:

> *Our senses give us only partial information about the objects present in our vicinity. In particular, vision gives only a* 2*-dimensional view of* 3*-dimensional objects. Our visual systems and our brains use sense information to learn about the* 3*-dimensional objects.*
>
> *Also humans and dogs can represent objects that are not presently visible.* (*The evidence about dogs is that if a thrown ball goes out of sight, the dog will look for it.*) *Humans can infer the existence of objects that are out of sight, and human learning from experience often involves learning about the hidden reality behind some phenomenon. This is what science is usually about, but it occurs in common sense reasoning as well.*
>
> *Machine learning research, to my knowledge, has so far involved classifying appearances, and has not involved inferring reality behind experience. Classifying experience is inadequate as a model of human learning and also inadequate for robotic applications.* [...] *Another way of looking at it is that we use observations to recognize patterns in the world, not just patterns in the observations.*

We propose next a learning model that makes explicit the dichotomy between appearance and reality, and respects the premises set forth above. We argue that such a model goes beyond simply being able to cope with missing information in learning examples. Instead, it shows that despite the use of supervised learning techniques, it is possible for such learning to be carried entirely autonomously — as is the case in scientific discovery — without the supervision of some external teacher. We call the new learning model ***autodidactic*** in recognition of this fact. We discuss other conceptual merits of autodidactic learning throughout this work.

2.1. Learning from partial observations

In the PAC learning model [28], an agent is given access to learning *examples* randomly drawn from an arbitrary, but fixed, probability distribution $\mathcal{D}$ over binary vectors $\mathtt{exm} \in \{0,1\}^{|\mathcal{A}|}$, for some fixed set of binary *attributes* $\mathcal{A}$. The examples are structured, in the sense that the value of a designated *target attribute* x_t is determined by some unknown, but fixed, function $\varphi \in \mathcal{C}$ of the remaining attributes $\mathcal{A} \setminus \{x_t\}$; the function φ is known as the *target concept*, and the class $\mathcal{C}$ of all possible target concepts is known as the *concept class*. During an initial learning phase an agent is expected to efficiently produce a *hypothesis* function $h \in \mathcal{H}$ that is highly accurate with respect to $\mathcal{D}$, in the sense that it predicts with high probability the value of the target attribute x_t in evaluation examples drawn from $\mathcal{D}$, given access to the values of only the remaining attributes $\mathcal{A} \setminus \{x_t\}$; the class $\mathcal{H}$ of all possible hypotheses is known as the *hypothesis class*.

Implicit in the definition of the PAC learning model is the premise that an agent has access to complete information on the values of attributes. Each example contains sufficient information to determine the value of the target attribute x_t; the primary challenge of the learning task, thus, is that of forming a hypothesis of *how* to determine the value of the target attribute *given* the values of the remaining attributes. In most realistic domains, however, an agent is burdened with an additional challenge: some of the information necessary to determine the value of the target attribute is missing in the examples. The agent has, therefore, access only to partial depictions of the learning examples, and this is the case during both the learning and the evaluation phase. These partial depictions of the learning examples we shall call ***observations***. Although in the general case observations could also be noisy with respect to the learning examples, we shall not consider such a scenario in this work, and we will henceforth assume that observations are only incomplete.

Observations are ternary vectors $\mathtt{obs} \in \{0,1,*\}^{|\mathcal{A}|}$, with the value $*$ indicating that the corresponding attribute is unobserved or "don't know". The mapping from examples to observations happens through a ***masking process***, a (stochastic) process $\mathtt{mask} : \{0,1\}^{|\mathcal{A}|} \rightarrow \{0,1,*\}^{|\mathcal{A}|}$ aimed to model an agent's sensors. The masking process $\mathtt{mask}$ induces a probability distribution $\mathtt{mask}(\mathtt{exm})$ over observations that may depend on the example $\mathtt{exm}$; we write $\mathtt{obs} \leftarrow \mathtt{mask}(\mathtt{exm})$ to denote that observation $\texttt{obs}$ is drawn from this probability distribution with non-zero probability. The noiseless nature of sensing amounts to insisting that whenever $\texttt{obs} \leftarrow \texttt{mask}(\texttt{exm})$ and for every attribute $x_i \in \mathcal{A}$, $\texttt{obs}[i] \in \{\texttt{exm}[i], *\}$, where $\texttt{obs}[i]$ and $\texttt{exm}[i]$ correspond, respectively, to the value of the i-th attribute according to $\texttt{obs}$ and $\texttt{exm}$. An observation $\texttt{obs}$ that is an image of an example $\texttt{exm}$ under *some* masking process is said to ***mask*** $\texttt{exm}$. Each attribute $x_i \in \mathcal{A}$ in an observation $\texttt{obs}$ with $\texttt{obs}[i] = *$ is said to be ***masked*** in $\texttt{obs}$.

Masking processes are, in general, many-to-many mappings from examples to observations. Given, for instance, the examples $\texttt{exm}_1 = \texttt{0010110}$ and $\texttt{exm}_2 = \texttt{0100100}$, and the observations $\texttt{obs}_1 = \texttt{0} * \texttt{10} * * \texttt{0}$, $\texttt{obs}_2 = * \texttt{10} * * \texttt{00}$, and $\texttt{obs}_3 = \texttt{0} * * \texttt{01} * *$, one masking process $\texttt{mask}$ is the following: on input $\texttt{exm}_1$ it returns $\texttt{obs}_1$ with probability 0.3 and $\texttt{obs}_3$ with probability 0.7; on input $\texttt{exm}_2$ it returns $\texttt{obs}_2$ with probability 0.6 and $\texttt{obs}_3$ with probability 0.4. The one-to-many nature of masking processes is intended to capture the stochastic nature of sensing. An agent attempting to sense the same reality twice (e.g., $\texttt{exm}_1$) may end up with two different appearances (e.g., $\texttt{obs}_1$ and $\texttt{obs}_3$). On the other hand, their many-to-one nature is intended to capture the loss of information due to an agent's limited sensing abilities. Two distinct realities (e.g., $\texttt{exm}_1$ and $\texttt{exm}_2$) may *appear* to be the same (e.g., $\texttt{obs}_3$) to an agent; no indication is given in the obtained appearance as to which reality was the one that was actually sensed.

The loss of information due to masking happens during *both* the learning *and* the evaluation phase. Thus, the agent never directly observes the learning examples, but has access only to observations that mask the learning examples. Yet, as in the PAC model, the agent is expected to produce a hypothesis for predicting the value of a target attribute. We emphasize that the hypothesis is a function over boolean attributes as is the case in the PAC model. In some sense, the agent is trying to encode in this hypothesis knowledge about the structure of the underlying examples — not knowledge about the structure of observations and the way the masking process hides information. Indeed, the central premise of this work is that information in observations is hidden in an arbitrary manner. The central question, then, is whether the structure of the underlying examples can still be learned in the PAC sense given such arbitrarily selected partial information. The PAC model can be viewed as the special case of our model when observations do not contain $*$ values.

To formalize the way that structure is present in examples (i.e., the requirement that the value of the target attribute is determined by some function of the remaining attributes), but also the way that predictions are made through a hypothesis function, we follow the PAC model and employ boolean formulas: syntactic objects over the set of attributes $\mathcal{A}$, associated with the typical semantics for evaluating them given a complete assignment of values to their attributes. Given a formula φ and an example $\texttt{exm}$, we write $\texttt{val}(\varphi \mid \texttt{exm})$ to denote the value of φ on $\texttt{exm}$. Unlike the PAC semantics, however, it is necessary to define also the value $\texttt{val}(\varphi \mid \texttt{obs})$ of a formula φ on an observation $\texttt{obs}$, since in the general case the agent will make predictions by evaluating a learned hypothesis on such a partial observation. Note that it is possible for $\texttt{val}(\varphi \mid \texttt{obs})$ to have a value even if $\texttt{obs}$ does not offer $\{\texttt{0}, \texttt{1}\}$ values for all the attributes in φ. On the other hand, if $\texttt{val}(\varphi \mid \texttt{obs})$ remains undetermined due to missing information, then we define $\texttt{val}(\varphi \mid \texttt{obs})$ to equal $*$, to indicate a "don't know" value for φ on $\texttt{obs}$.

Note that evaluating some formula on some observation cannot necessarily be done efficiently, even if the formula is efficiently evaluatable on every example. Indeed, evaluating 3-CNF formulas on the observation in which all attributes are masked is as hard as deciding whether 3-CNF formulas have a satisfying assignment; an $\texttt{NP}$-complete problem [8]. Most definitions and results that we later state do not rely on actually evaluating formulas (efficiently); hence, they are not conditioned on the formulas involved being efficiently evaluatable. When formulas need to be efficiently evaluatable for a result to hold, this is stated explicitly in the conditions of the result. It remains open whether formulas that are not efficiently evaluatable on observations, but are so on examples, can be learned in the sense defined later on.

We proceed now to define how (a particular type of) structure is encoded in examples.

Definition 2.1 *(Supported concept classes).* A target attribute $x_t \in \mathcal{A}$ is ***expressed by*** a formula φ over $\mathcal{A} \setminus \{x_t\}$ ***w.r.t.*** a probability distribution $\mathcal{D}$ if

$$Pr\big[\texttt{val}(\varphi \mid \texttt{exm}) = \texttt{exm}[t] \mid \texttt{exm} \leftarrow \mathcal{D}\big] = 1.$$

A probability distribution $\mathcal{D}$ ***supports*** a concept class $\mathcal{C}$ of formulas over $\mathcal{A} \setminus \{x_t\}$ ***for*** a target attribute $x_t \in \mathcal{A}$ if there exists a formula $c \in \mathcal{C}$ such that x_t is expressed by c w.r.t. $\mathcal{D}$; c is the ***target concept for*** x_t ***under*** $\mathcal{D}$.

We view the values of *all* attributes as being drawn from some probability distribution $\mathcal{D}$. We regard this approach as corresponding more closely to what conceptually happens in certain domains, than the approach typically taken by supervised learning models: The attributes that encode the state of affairs are not a priori distinguished into target and non-target attributes; they are all equivalent, and nature, as captured by the probability distribution $\mathcal{D}$, assigns a value to each of these attributes. An agent's sensors may then mask some of the attributes without distinguishing any one of them. The distinction of a target attribute, and the assumption that this attribute is somehow correlated with the rest of the attributes serve only as premises of a particular type of learning task, and such a correlation is imposed by appropriately restricting $\mathcal{D}$. Indeed, under this view it is possible to easily generalize Definition 2.1 to encode other types of correlation between attributes, like, for instance, that the number of attributes in $\mathcal{A}$ that are assigned the value $\texttt{1}$ in an example $\texttt{exm} \leftarrow \mathcal{D}$ is divisible by 3 with probability 0.95. Although we find these types of correlation intriguing, and meriting further investigation, we focus in this work on the type of correlation stated in Definition 2.1.

To complete the description of our learning model, we need to define when a prediction of a formula is consistent with respect to an observation. In other words, we wish to determine when the two sources of information available to an agent, its sensor readings, and the conclusions it draws through (learned) rules, agree with each other. We define a formula φ to have a ***consistency conflict with*** a target attribute x_t ***w.r.t.*** an observation obs if $\{\mathtt{val}(\varphi \mid \mathtt{obs}), \mathtt{obs}[t]\} = \{\mathtt{0}, \mathtt{1}\}$; the value of the formula and the *observed* value of the target attribute are both $\{\mathtt{0}, \mathtt{1}\}$, and differ from each other. In all other cases, either the two suggested $\{\mathtt{0}, \mathtt{1}\}$ values agree, or at least one of the two suggested values is "don't know". We extend the notion of consistency to apply to a probability distribution over observations. This probability distribution we denote by $\mathtt{mask}(\mathcal{D})$ to indicate that it is induced by the probability distribution $\mathcal{D}$ from which examples are drawn, and the masking process mask through which these examples are mapped to observations.

Definition 2.2 *(Degree of consistency).* A formula φ over $\mathcal{A} \setminus \{x_t\}$ is $(1-\varepsilon)$-***consistent with*** a target attribute $x_t \in \mathcal{A}$ ***under*** a probability distribution $\mathcal{D}$ ***and*** a masking process mask if

$$Pr\big[\{\mathtt{val}(\varphi \mid \mathtt{obs}), \mathtt{obs}[t]\} = \{\mathtt{0}, \mathtt{1}\} \mid \mathtt{exm} \leftarrow \mathcal{D}; \mathtt{obs} \leftarrow \mathtt{mask}(\mathtt{exm})\big] \leqslant \varepsilon.$$

We now state formally the learning requirements under the autodidactic learning model that we consider. In what follows we will denote a ***learning task over*** $\mathcal{A}$ by a triple $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$, where x_t is a target attribute in $\mathcal{A}$, $\mathcal{C}$ is a concept class of formulas over $\mathcal{A} \setminus \{x_t\}$, and $\mathcal{H}$ a hypothesis class of formulas over $\mathcal{A} \setminus \{x_t\}$.

Definition 2.3 *(Consistent learnability).* An algorithm $\mathcal{L}$ is a ***consistent learner for*** a learning task $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ over $\mathcal{A}$ if for every probability distribution $\mathcal{D}$ supporting $\mathcal{C}$ for x_t, every masking process mask, every real number $\delta \in (0, 1]$, and every real number $\varepsilon \in (0, 1]$, algorithm $\mathcal{L}$ has the following property: given access to $\mathcal{A}$, $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$, δ, ε, and an oracle returning observations drawn from $\mathtt{mask}(\mathcal{D})$, algorithm $\mathcal{L}$ runs in time polynomial in $1/\delta$, $1/\varepsilon$, $|\mathcal{A}|$, and the size of the target concept for x_t under $\mathcal{D}$, and returns, with probability $1-\delta$, a hypothesis $h \in \mathcal{H}$ that is $(1-\varepsilon)$-consistent with x_t under $\mathcal{D}$ and mask. The concept class $\mathcal{C}$ over $\mathcal{A} \setminus \{x_t\}$ is ***consistently learnable on*** the target attribute $x_t \in \mathcal{A}$ ***by*** the hypothesis class $\mathcal{H}$ over $\mathcal{A} \setminus \{x_t\}$ if there exists a consistent learner for $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ over $\mathcal{A}$.

The definition of consistent learnability follows closely the PAC semantics, with the added requirement that learnability succeeds for an arbitrary masking process mask, and not only when mask is the identity mapping (as is the case under the PAC semantics). Although the added requirement might at first seem too arduous, recall that exactly in those situations where learnability becomes harder due to missing information, formulas may make "don't know" predictions more freely, avoiding thus consistency conflicts. We emphasize, however, that "don't know" predictions cannot be abused, since a learner may not produce a hypothesis that arbitrarily chooses to abstain from making predictions. It is the masking process that gives a formula the ability to make "don't know" predictions, and this is beyond the control of the learner. We later contrast our approach to other models of learning where hypotheses *actively choose* when to abstain from making predictions. In such models, one is required to introduce a second metric for measuring success of a hypothesis: its degree of completeness; the probability with which a $\{\mathtt{0}, \mathtt{1}\}$ prediction is made.

2.2. Are accurate predictions possible?

With a complete proposal for a learning model, we now revisit our original motivation for developing a model for learning from partial observations: to recover missing information in the incomplete sensory inputs of an agent. Does our definition of consistent learnability address this goal? Recall that a highly consistent formula is guaranteed to make predictions that are consistent with randomly drawn observations. In particular, in those cases where the problem of missing information is interesting, namely when the target attribute is masked in an observation, the consistency guarantee seems to offer essentially nothing, since any prediction is consistent with such an observation. What we need, therefore, is a notion of predictive correctness, not with respect to the observations, but with respect to the underlying examples: an agent wishes to be able to match the *unobserved* reality behind the appearances of its environment.

We define a formula φ to have an ***accuracy conflict with*** a target attribute x_t ***w.r.t.*** an observation obs ***obtained from*** an example exm if $\{\mathtt{val}(\varphi \mid \mathtt{obs}), \mathtt{exm}[t]\} = \{\mathtt{0}, \mathtt{1}\}$; the value of the formula and the *actual* value of the target attribute are both $\{\mathtt{0}, \mathtt{1}\}$, and differ from each other. In all other cases, either the two suggested $\{\mathtt{0}, \mathtt{1}\}$ values agree, or the value of the formula is "don't know". As in the case of consistency, we extend the notion of accuracy to apply to a probability distribution over observations.

Definition 2.4 *(Degree of accuracy).* A formula φ over $\mathcal{A} \setminus \{x_t\}$ is $(1-\varepsilon)$-***accurate w.r.t.*** a target attribute $x_t \in \mathcal{A}$ ***under*** a probability distribution $\mathcal{D}$ ***and*** a masking process mask if

$$Pr\big[\{\mathtt{val}(\varphi \mid \mathtt{obs}), \mathtt{exm}[t]\} = \{\mathtt{0}, \mathtt{1}\} \mid \mathtt{exm} \leftarrow \mathcal{D}; \mathtt{obs} \leftarrow \mathtt{mask}(\mathtt{exm})\big] \leqslant \varepsilon.$$

It is now natural to ask that our definition of learnability be revised so that highly accurate (instead of highly consistent) hypotheses be returned. A naive revision would, however, lead to a vacuous definition, where learnability would be trivially unattainable. Indeed, when the masking process is such that the target attribute and only the target attribute is masked in all observations, then clearly the learning algorithm has no access to any information about the value of the target attribute. Yet, any formula over the remaining attributes will be forced to make a $\{\mathtt{0},\mathtt{1}\}$ prediction. It is, then, impossible to determine which one amongst two formulas has a higher degree of accuracy, as both formulas will always make $\{\mathtt{0},\mathtt{1}\}$ predictions, but no feedback will be provided as to which, if any, of the two formulas makes a correct prediction. This compromises learnability even in domains where the concept and hypothesis classes contain only two formulas.

Theorem 2.1 *(Statistical indistinguishability in adversarial settings). Consider a target attribute $x_t \in \mathcal{A}$, a class $\mathcal{F}$ of formulas over $\mathcal{A} \setminus \{x_t\}$, and two formulas $\varphi_1, \varphi_2 \in \mathcal{F}$ such that $\varphi_1 \notin \{\varphi_2, \overline{\varphi_2}\}$. For every real number $\varepsilon \in [0, 1]$, there exist probability distributions $\mathcal{D}_1, \mathcal{D}_2$, and a masking process $\mathtt{mask}_0$, such that:*

(i) *φ_1 is 1-accurate and φ_2 is not more than $(1-\varepsilon)$-accurate, both w.r.t. x_t under $\mathcal{D}_1$ and $\mathtt{mask}_0$;*
(ii) *φ_1 is not more than $(1-\varepsilon)$-accurate and φ_2 is 1-accurate, both w.r.t. x_t under $\mathcal{D}_2$ and $\mathtt{mask}_0$;*
(iii) *$\mathtt{mask}_0(\mathcal{D}_1) = \mathtt{mask}_0(\mathcal{D}_2)$, and no attribute in $\mathcal{A} \setminus \{x_t\}$ is masked in any drawn observation.*

Proof. Let S be the set of truth-assignments to the attributes $\mathcal{A} \setminus \{x_t\}$ for which φ_1, φ_2 are assigned different truth-values. Fix any probability distribution $\mathcal{D}$ over all truth-assignments to the attributes $\mathcal{A} \setminus \{x_t\}$ that assigns probability ε to the set S, and probability $1-\varepsilon$ to the complement of S; since $\varphi_1 \notin \{\varphi_2, \overline{\varphi_2}\}$, both S and its complement are non-empty. For each $i \in \{1, 2\}$, extend $\mathcal{D}$ to the probability distribution $\mathcal{D}_i$ over examples from $\{\mathtt{0},\mathtt{1}\}^{|\mathcal{A}|}$, by completing the truth-assignments to the attributes $\mathcal{A} \setminus \{x_t\}$ so as to assign the induced truth-value of φ_i to the target attribute x_t. Choose $\mathtt{mask}_0$ to be the masking process that maps each example $\mathtt{exm}$ to an observation $\mathtt{obs}$ in which x_t is masked if and only if $\mathtt{val}(\varphi_1 \mid \mathtt{exm}) \neq \mathtt{val}(\varphi_2 \mid \mathtt{exm})$, and no attribute in $\mathcal{A} \setminus \{x_t\}$ is masked. By construction of $\mathtt{mask}_0$, for each example $\mathtt{exm}$ and each observation $\mathtt{obs} \leftarrow \mathtt{mask}_0(\mathtt{exm})$, it holds that x_t is masked in $\mathtt{obs}$ if and only if $\mathtt{val}(\varphi_1 \mid \mathtt{obs}) \neq \mathtt{val}(\varphi_2 \mid \mathtt{obs})$. By construction of $\mathcal{D}_1, \mathcal{D}_2$, and $\mathtt{mask}_0$, all three conditions of the claim follow. □

It becomes evident that although masking processes may arbitrarily hide information, completely ignoring the extent to which information is hidden may prevent us from attaining a meaningful and useful notion of learnability. Given a moment's thought, this is a natural conclusion. Structure exists in examples, yet a learner attempts to learn this structure given access only to observations. Thus, learnability becomes possible only if the masking process allows some of the structure of the underlying examples to carry over to the observations. In other words, we expect the observations to occasionally provide some feedback, according to the structure of the underlying examples, as to whether a candidate hypothesis is indeed highly accurate. The extent to which such feedback is provided depends on the masking process, and is quantified next.

Feedback is necessary only when a candidate hypothesis errs, i.e., has an accuracy conflict with the target attribute. Recall that an agent is not necessarily aware of an accuracy conflict. By way of illustration, if the target attribute x_3 is masked in an observation $\mathtt{10{*}1010}$, then φ has an accuracy conflict with the observation depending on which of the examples $\mathtt{1001010}, \mathtt{1011010}$ the observation was obtained from; the agent is oblivious to the choice of example, and hence to the existence of an accuracy conflict. In case of such accuracy conflicts, we expect that with some probability the value of the target attribute will be made known to the agent, so that the conflict might be detected. That is, we expect that any particular reality will not be indefinitely sensed by an agent without the agent realizing that it is making a wrong prediction.

Definition 2.5 *(Degree of concealment).* A masking process $\mathtt{mask}$ is $(1-\eta)$***-concealing for*** a target attribute $x_t \in \mathcal{A}$ ***w.r.t.*** a class $\mathcal{F}$ of formulas over $\mathcal{A} \setminus \{x_t\}$ if $\eta \in [0, 1]$ is the minimum value of

$$Pr\big[\mathtt{obs}[t] \neq * \mid \mathtt{obs} \leftarrow \mathtt{mask}(\mathtt{exm});\ \big\{\mathtt{val}(\varphi \mid \mathtt{obs}), \mathtt{exm}[t]\big\} = \{\mathtt{0},\mathtt{1}\}\big]$$

across all choices of an example $\mathtt{exm} \in \{\mathtt{0},\mathtt{1}\}^{|\mathcal{A}|}$, and a formula $\varphi \in \mathcal{F}$.[1]

The concealment degree in Definition 2.5 is a *worst-case* bound, across all possible examples, and all possible formulas used for making predictions. Each pair of a formula φ and an example $\mathtt{exm}$ imposes a constraint on η, and implies a lower bound on the concealment degree $1-\eta$ of the masking process $\mathtt{mask}$. Note that a probability distribution $\mathcal{D}$ may assign zero probability to the examples that "bring out" the adversarial nature of a masking process, making it, thus, look less adversarial than in the worst case. Note, also, that the concealment degree of a masking process may vary arbitrarily across target attributes.

As an illustration, consider a particular domain in which the target attribute is x_6, one of the formulas is $\varphi = (x_2 \vee \overline{x_4}) \oplus x_7$, where $\oplus$ denotes the "exclusive or" binary operator, and one of the examples is $\mathtt{exm} = \mathtt{0110011}$. The masking process

[1] A conditional probability is undefined when the event in its condition occurs with probability 0. In such cases, we define the conditional probability to equal 1. In the context of Definition 2.5, this choice implies that cases in which a formula does not have an accuracy conflict with the target attribute w.r.t. any of the observations obtained from a particular example, can be safely ignored, since such cases do not constrain the concealment degree of the masking process in any way.

Table 1
Observations obtained from the example $\texttt{exm} = \texttt{0110011}$ by applying a particular masking process $\texttt{mask}$, and the corresponding predictions of formula $\varphi = (x_2 \vee \overline{x_4}) \oplus x_7$ for the target attribute x_6.

$\texttt{obs} \leftarrow \texttt{mask(exm)}$		Prediction for x_6	
observation	probability	$\texttt{val}(\varphi \mid \texttt{obs})$	accuracy conflict
$0 * 1\ 0 * 1\ 1$	0.27	0	yes
$* 1 * 0 * 1\ 1$	0.15	0	yes
$* 1\ 1 * 0 * 1$	0.33	0	yes
$0 * 1\ 0 * 1 *$	0.04	$*$	no
$0 * 1 * 0 * 1$	0.21	$*$	no

$\texttt{mask}$ is such that $\texttt{exm}$ is mapped to observations as shown in Table 1. According to $\texttt{mask}$, the observations that give rise to accuracy conflicts are drawn with a total probability of 0.75. Among those, the observations in which the target attribute x_6 is not masked are drawn with a total probability of 0.42. By the law of conditional probabilities, it follows that

$$Pr\big[\texttt{obs}[t] \neq * \mid \texttt{obs} \leftarrow \texttt{mask(exm)};\ \big\{\texttt{val}(\varphi \mid \texttt{obs}), \texttt{exm}[t]\big\} = \{\texttt{0}, \texttt{1}\}\big] = \frac{0.42}{0.75} = 0.56.$$

Definition 2.5 now implies that $\eta \leqslant 0.56$, and the masking process $\texttt{mask}$ is at least 0.44-concealing. Additional formulas and examples may impose extra bounds on η, which, in turn, may increase the concealment degree of $\texttt{mask}$. If none of the extra bounds on η is smaller than 0.56, then the definition of η would imply that the masking process $\texttt{mask}$ is exactly 0.44-concealing for the particular target attribute x_6.

2.3. Going from consistency to accuracy

Given a crisp metric of the degree of feedback that a masking process provides to an agent, it is easy to see that the negative learnability result of Theorem 2.1 holds precisely because it appeals to a 1-concealing masking process. As in other learning models were some parameterized resource renders learnability impossible when the parameter riches some relevant limit (e.g., in the case of random classification noise [1], learnability becomes impossible when the noise rate becomes $1/2$), we could extend the definition of learnability to allow resources that grow inversely with the distance of this parameter from its limit. In the case of the concealment degree the limit is 1, and η defines the distance of the concealment degree of a $(1-\eta)$-concealing masking process $\texttt{mask}$ from this limit. Hence, we could revise the definition of learnability so that highly accurate (instead of highly consistent) hypotheses be returned, but at the same time allow additional resources that grow with $1/\eta$ to account for the adversarial nature with which $\texttt{mask}$ may hide information. A learner expected to return a $(1-\varepsilon)$-accurate hypothesis could then exploit the additional resources in order to obtain a $(1-\eta \cdot \varepsilon)$-consistent hypothesis, and then appeal to the following result to establish that this hypothesis is, in fact, $(1-\varepsilon)$-accurate. The proof of the next result builds on the natural realization that a prediction is consistent if and only if it is either accurate or the target attribute is masked. Informally, then, in set-theoretic terms it holds that *consistency* = *accuracy* ∪ *concealment*.

Theorem 2.2 *(The relation of consistency and accuracy). Consider a target attribute $x_t \in \mathcal{A}$, and a class $\mathcal{F}$ of formulas over $\mathcal{A} \setminus \{x_t\}$. For every real number $\eta \in [0,1]$, and every masking process* $\texttt{mask}$ *that is $(1-\eta)$-concealing for x_t w.r.t. $\mathcal{F}$, the following conditions hold*:

(i) *for every probability distribution $\mathcal{D}$, and every formula $\varphi \in \mathcal{F}$, it holds that*: *φ is $(1-\varepsilon)$-accurate w.r.t. x_t under $\mathcal{D}$ and* $\texttt{mask}$ *for some real number $\varepsilon \in [0,1]$, if φ is $(1-\eta \cdot \varepsilon)$-consistent with x_t under $\mathcal{D}$ and* $\texttt{mask}$, *and $\eta \neq 0$*;
(ii) *there exists a probability distribution $\mathcal{D}_0$, and a formula $\varphi_0 \in \mathcal{F}$, such that*: *$\varphi_0$ is $(1-\varepsilon)$-accurate w.r.t. x_t under $\mathcal{D}_0$ and* $\texttt{mask}$ *for some real number $\varepsilon \in [0,1]$, only if φ_0 is $(1-\eta \cdot \varepsilon)$-consistent with x_t under $\mathcal{D}_0$ and* $\texttt{mask}$.

Proof. For every formula $\varphi \in \mathcal{F}$, every example $\texttt{exm} \in \{\texttt{0},\texttt{1}\}^{|\mathcal{A}|}$, every observation $\texttt{obs} \in \{\texttt{0},\texttt{1},*\}^{|\mathcal{A}|}$ that masks $\texttt{exm}$, and every probability distribution $\mathcal{D}$, denote by $\mathbb{E}_{\text{DR}}(\texttt{exm}, \texttt{obs}, \mathcal{D})$ the event that $\texttt{exm} \leftarrow \mathcal{D}$ and $\texttt{obs} \leftarrow \texttt{mask(exm)}$, by $\mathbb{E}_{\text{CC}}(\varphi, \texttt{obs})$ the event that φ has a consistency conflict with x_t w.r.t. $\texttt{obs}$, by $\mathbb{E}_{\text{AC}}(\varphi, \texttt{exm}, \texttt{obs})$ the event that φ has an accuracy conflict with x_t w.r.t. $\texttt{obs}$ obtained from $\texttt{exm}$, and by $\mathbb{E}_{\text{NM}}(\texttt{obs})$ the event that x_t is not masked in $\texttt{obs}$. Clearly, the event $\mathbb{E}_{\text{CC}}(\varphi, \texttt{obs})$ holds exactly when the events $\mathbb{E}_{\text{AC}}(\varphi, \texttt{exm}, \texttt{obs})$ and $\mathbb{E}_{\text{NM}}(\texttt{obs})$ hold simultaneously. In particular, this is true even when $\texttt{exm}$ and $\texttt{obs}$ are restricted so that the event $\mathbb{E}_{\text{DR}}(\texttt{exm}, \texttt{obs}, \mathcal{D})$ is true. Thus,

$$Pr\big[\mathbb{E}_{\text{CC}}(\varphi, \texttt{obs}) \mid \mathbb{E}_{\text{DR}}(\texttt{exm}, \texttt{obs}, \mathcal{D})\big] = Pr\big[\mathbb{E}_{\text{AC}}(\varphi, \texttt{exm}, \texttt{obs}) \wedge \mathbb{E}_{\text{NM}}(\texttt{obs}) \mid \mathbb{E}_{\text{DR}}(\texttt{exm}, \texttt{obs}, \mathcal{D})\big].$$

From the law of conditional probabilities, the right hand side of the equation equals

$$Pr\big[\mathbb{E}_{\text{NM}}(\texttt{obs}) \mid \mathbb{E}_{\text{DR}}(\texttt{exm}, \texttt{obs}, \mathcal{D}) \wedge \mathbb{E}_{\text{AC}}(\varphi, \texttt{exm}, \texttt{obs})\big] \cdot Pr\big[\mathbb{E}_{\text{AC}}(\varphi, \texttt{exm}, \texttt{obs}) \mid \mathbb{E}_{\text{DR}}(\texttt{exm}, \texttt{obs}, \mathcal{D})\big].$$

We proceed to derive bounds for the first term of the product above. In the first direction, Definition 2.5 implies that for every example $\texttt{exm} \in \{\texttt{0},\texttt{1}\}^{|\mathcal{A}|}$, and every formula $\varphi \in \mathcal{F}$, it holds that:

$$Pr\big[\mathtt{obs}[t] \neq * \mid \mathtt{obs} \leftarrow \mathtt{mask}(\mathtt{exm}); \big\{\mathtt{val}(\varphi \mid \mathtt{obs}), \mathtt{exm}[t]\big\} = \{0,1\}\big] \geqslant \eta.$$

Now, if $\mathtt{exm}$ is drawn from any given probability distribution $\mathcal{D}$, the overall probability that $\mathtt{obs}[t] \neq *$ given that $\{\mathtt{val}(\varphi \mid \mathtt{obs}), \mathtt{exm}[t]\} = \{0,1\}$ remains lower bounded by η. Thus,

$$Pr\big[\mathbb{E}_{\mathrm{NM}}(\mathtt{obs}) \mid \mathbb{E}_{\mathrm{DR}}(\mathtt{exm}, \mathtt{obs}, \mathcal{D}) \wedge \mathbb{E}_{\mathrm{AC}}(\varphi, \mathtt{exm}, \mathtt{obs})\big] \geqslant \eta. \tag{1}$$

In the other direction, Definition 2.5 implies that there exists an example $\mathtt{exm}_0 \in \{0,1\}^{|\mathcal{A}|}$, and a formula $\varphi_0 \in \mathcal{F}$, such that:

$$Pr\big[\mathtt{obs}[t] \neq * \mid \mathtt{obs} \leftarrow \mathtt{mask}(\mathtt{exm}_0); \big\{\mathtt{val}(\varphi_0 \mid \mathtt{obs}), \mathtt{exm}_0[t]\big\} = \{0,1\}\big] \leqslant \eta.$$

Now, if $\mathtt{exm}_0$ is replaced with an example $\mathtt{exm}$ drawn from the probability distribution $\mathcal{D}_0$ that is defined so that it assigns probability 1 to $\mathtt{exm}_0$ being drawn, the probability that $\mathtt{obs}[t] \neq *$ given that $\{\mathtt{val}(\varphi_0 \mid \mathtt{obs}), \mathtt{exm}[t]\} = \{0,1\}$ remains upper bounded by η. Thus,

$$Pr\big[\mathbb{E}_{\mathrm{NM}}(\mathtt{obs}) \mid \mathbb{E}_{\mathrm{DR}}(\mathtt{exm}, \mathtt{obs}, \mathcal{D}_0) \wedge \mathbb{E}_{\mathrm{AC}}(\varphi_0, \mathtt{exm}, \mathtt{obs})\big] \leqslant \eta. \tag{2}$$

Finally, we proceed to establish that the conditions of the claim hold. For Condition (i), fix an arbitrary probability distribution $\mathcal{D}$, and an arbitrary formula $\varphi \in \mathcal{F}$, and assume that φ is $(1 - \eta \cdot \varepsilon)$-consistent with x_t under $\mathcal{D}$ and $\mathtt{mask}$ for some real number $\varepsilon \in [0,1]$, and that $\eta \neq 0$. Then, $Pr[\mathbb{E}_{\mathrm{CC}}(\varphi, \mathtt{obs}) \mid \mathbb{E}_{\mathrm{DR}}(\mathtt{exm}, \mathtt{obs}, \mathcal{D})] \leqslant \eta \cdot \varepsilon$, or equivalently

$$Pr\big[\mathbb{E}_{\mathrm{NM}}(\mathtt{obs}) \mid \mathbb{E}_{\mathrm{DR}}(\mathtt{exm}, \mathtt{obs}, \mathcal{D}) \wedge \mathbb{E}_{\mathrm{AC}}(\varphi, \mathtt{exm}, \mathtt{obs})\big] \cdot Pr\big[\mathbb{E}_{\mathrm{AC}}(\varphi, \mathtt{exm}, \mathtt{obs}) \mid \mathbb{E}_{\mathrm{DR}}(\mathtt{exm}, \mathtt{obs}, \mathcal{D})\big] \leqslant \eta \cdot \varepsilon.$$

Since $\eta \neq 0$, Inequality (1) immediately implies that $Pr[\mathbb{E}_{\mathrm{AC}}(\varphi, \mathtt{exm}, \mathtt{obs}) \mid \mathbb{E}_{\mathrm{DR}}(\mathtt{exm}, \mathtt{obs}, \mathcal{D})] \leqslant \varepsilon$. Therefore φ is $(1 - \varepsilon)$-accurate w.r.t. x_t under $\mathcal{D}$ and $\mathtt{mask}$.

For Condition (ii), consider the probability distribution $\mathcal{D}_0$, and the formula $\varphi_0 \in \mathcal{F}$, both as defined in the context of Inequality (2), and assume that φ_0 is $(1 - \varepsilon)$-accurate w.r.t. x_t under $\mathcal{D}_0$ and $\mathtt{mask}$ for some real number $\varepsilon \in [0,1]$. Then, $Pr[\mathbb{E}_{\mathrm{AC}}(\varphi_0, \mathtt{exm}, \mathtt{obs}) \mid \mathbb{E}_{\mathrm{DR}}(\mathtt{exm}, \mathtt{obs}, \mathcal{D}_0)] \leqslant \varepsilon$. Inequality (2) immediately implies that

$$Pr\big[\mathbb{E}_{\mathrm{NM}}(\mathtt{obs}) \mid \mathbb{E}_{\mathrm{DR}}(\mathtt{exm}, \mathtt{obs}, \mathcal{D}_0) \wedge \mathbb{E}_{\mathrm{AC}}(\varphi_0, \mathtt{exm}, \mathtt{obs})\big] \cdot Pr\big[\mathbb{E}_{\mathrm{AC}}(\varphi_0, \mathtt{exm}, \mathtt{obs}) \mid \mathbb{E}_{\mathrm{DR}}(\mathtt{exm}, \mathtt{obs}, \mathcal{D}_0)\big] \leqslant \eta \cdot \varepsilon,$$

or equivalently $Pr[\mathbb{E}_{\mathrm{CC}}(\varphi_0, \mathtt{obs}) \mid \mathbb{E}_{\mathrm{DR}}(\mathtt{exm}, \mathtt{obs}, \mathcal{D}_0)] \leqslant \eta \cdot \varepsilon$. Therefore φ_0 is $(1 - \eta \cdot \varepsilon)$-consistent with x_t under $\mathcal{D}_0$ and $\mathtt{mask}$. The claim follows. □

Condition (i) of Theorem 2.2 provides a formal implication from highly consistent hypotheses to highly accurate ones. There is, however, a caveat to this implication. The degree of accuracy of the predictions is not necessarily as high as their degree of consistency. Given a moment's thought, this makes perfect sense. The requirement of making accurate predictions is a stronger one as compared to that of making consistent predictions. In those cases that the target attribute is not masked, accuracy conflicts and consistency conflicts are equivalent, whereas in those cases that the target attribute is masked, consistency conflicts never occur, while accuracy conflicts are still possible. What is perhaps more intriguing is the fact that the extent to which the degree of accuracy diminishes with respect to the degree of consistency, depends on the degree of concealment of the masking process. Assuming that our sensors and physical world do not adversarially hide information from us, one may interpret the above result as corroborating that the approach humans follow in using consistent theories for recovering missing information in their appearances is a rational strategy.

The dependence of accuracy on the concealment degree explains also why it is possible in certain cases for the implication from consistency to accuracy to be violated. This happens exactly in those cases where the concealment degree is high, and thus η is close to zero. Condition (ii) of Theorem 2.2 establishes that there exist domains in which certain formulas with a high degree of consistency have a low degree of accuracy. At the same time, Condition (ii) suggests also that the use of consistent hypotheses is, in the worst case, an optimal strategy for recovering missing information; in certain domains, the bound on the degree of accuracy that Condition (i) guarantees for a highly consistent formula, is tight. Furthermore, this optimality is guaranteed without any knowledge of the concealment degree, which, as we will later discuss in Section 3.2, may be hard or even impossible to determine, for all but very simple masking processes.

Obtaining highly accurate hypotheses through highly consistent hypotheses (cf. Definition 2.3) is, thus, a valid and optimal, in the worst case, approach. Still, why should a direct approach of learning highly accurate hypotheses not be used instead? The answer is simple. Consistency is a much more natural notion to work with, and avoids complications arising from having to deal with the degree of concealment of a masking process. More importantly, a formula's degree of consistency can be *reliably empirically estimated* as the following simple result shows, while its degree of accuracy cannot (cf. Theorem 2.1), as that would require access to the value of the target attribute even when the target attribute is masked in observations.

Definition 2.6 *(Degree of consistency (sample version)).* A formula φ over $\mathcal{A} \setminus \{x_t\}$ is $(1 - \varepsilon)$***-consistent with*** a target attribute $x_t \in \mathcal{A}$ ***given*** a sample $\mathcal{O}$ of observations if φ has a consistency conflict with x_t w.r.t. at most an ε fraction of the observations in $\mathcal{O}$.

Theorem 2.3 *(Empirical estimability of consistency degree). Consider a target attribute $x_t \in \mathcal{A}$, a formula φ over $\mathcal{A} \setminus \{x_t\}$, a probability distribution $\mathcal{D}$, a masking process* mask*, and a sample $\mathcal{O}$ of observations drawn independently from* $\mathtt{mask}(\mathcal{D})$*. For every pair of real numbers $\varepsilon, \gamma \in [0, 1]$, if φ is $(1-\varepsilon)$-consistent with x_t given $\mathcal{O}$, then, with probability at least $1 - e^{-2|\mathcal{O}|\gamma^2}$, it holds that φ is $(1-(\varepsilon+\gamma))$-consistent with x_t under $\mathcal{D}$ and* mask*.*

Proof. For each observation $\mathtt{obs}_i \in \mathcal{O}$, let the random variable X_i be the indicator variable for the event $\mathbb{E}_{\mathrm{cc}}(\varphi, \mathtt{obs}_i)$ that φ has a consistency conflict with x_t w.r.t. $\mathtt{obs}_i$; by construction of $\mathcal{O}$ the random variables are independent. Define the random variable $X \triangleq |\mathcal{O}|^{-1} \cdot \sum_{i=1}^{|\mathcal{O}|} X_i$ to be the mean of these random variables. By linearity of expectations, $E[X]$ is the mean $|\mathcal{O}|^{-1} \cdot \sum_{i=1}^{|\mathcal{O}|} E[X_i]$ of the expectations of these random variables. By standard Hoeffding concentration bounds [12], the probability that $|X - E[X]| > \gamma$ is at most $e^{-2|\mathcal{O}|\gamma^2}$. Clearly, X is the fraction of the observations in $\mathcal{O}$ w.r.t. which φ has a consistency conflict with x_t, and therefore $X \leqslant \varepsilon$. By definition of the random variables, $E[X] = E[X_1] = Pr[\mathbb{E}_{\mathrm{cc}}(\varphi, \mathtt{obs}_1)]$; thus φ is $(1 - E[X])$-consistent with x_t under $\mathcal{D}$ and mask. Since $X \leqslant \varepsilon$, it follows that with probability $1 - e^{-2|\mathcal{O}|\gamma^2}$, $E[X] \leqslant \varepsilon + \gamma$, as needed. □

Whether a formula's degree of consistency can be reliably empirically estimated in an *efficient* manner is an orthogonal issue, and depends on whether the formula can be evaluated efficiently on partial observations.

3. Discussion and related work

The problem of missing information in learning settings had been recognized early on in the literature. Valiant [28] himself in the paper that introduced PAC learning had, in fact, considered some form of learning from partial observations. Various frameworks developed since then have offered solutions to related problems. Within the Learning Theory community, extensions of the PAC model have been proposed to deal, to varying degrees, with the problem of missing information. Within the broader Machine Learning community the problem of dealing with missing information has received significant attention, especially in devising practical solutions in real-world settings. Other communities within the area of Artificial Intelligence and within Computer Science at large have also offered solutions to problems related to the manipulation of incomplete data. Fields outside Computer Science have also dealt with the problem, especially within the area of Statistical Analysis. It is beyond the scope of this work to do a full survey of the problems that have been examined and the solutions that have been offered. In this section we will mostly focus on discussing work closely related to ours: extensions of the PAC model that deal with missing information.

In the sequel we defend certain modelling choices within our framework. We then contrast the degree of concealment as a measure of the degree of missing information to other standard metrics found in the Statistical Analysis literature. We finally consider related PAC learning frameworks and discuss how those relate to autodidactic learning. We identify three dimensions along which these frameworks may be compared and contrasted to each other and to autodidactic learning: (i) the semantics of "don't know" predictions, (ii) the degree of supervision while learning, and (iii) the regularity on how information is hidden in observations.

3.1. Are "don't know" predictions justified?

Our choices as to when "don't know" predictions are considered justified, and as to how such predictions are accounted for when measuring a formula's degree of accuracy (cf. Definition 2.4), may raise certain objections. We discuss here three classes of objections that we have identified.

A first objection relates to our choice of when "don't know" predictions are allowed. Recall that a formula φ predicts $*$ on an observation $\mathtt{obs}_0$ if and only if $\mathtt{val}(\varphi \mid \mathtt{obs}_0)$ is undetermined. It could be argued that a "don't know" prediction might not be justified if, for instance, φ evaluates to $\mathtt{1}$ on the vast majority (but not all, since $\mathtt{val}(\varphi \mid \mathtt{obs}_0)$ is undefined) of the examples masked by $\mathtt{obs}_0$; would it not be more reasonable to define the value of φ on $\mathtt{obs}_0$ to be $\mathtt{1}$ in this case? Our answer is no. Just because the majority of the *possible* underlying examples $\mathtt{exm}$ of observation $\mathtt{obs}_0$ are such that $\mathtt{val}(\varphi \mid \mathtt{exm}) = \mathtt{1}$, it does not follow that such examples will be drawn with high, or even non-zero, probability from the underlying probability distribution $\mathcal{D}$. That is, there is no way to exclude the eventuality that the agent's environment will supply the agent only with examples $\mathtt{exm}$ such that $\mathtt{val}(\varphi \mid \mathtt{exm}) = \mathtt{0}$, which would then completely undermine the reason for choosing to define the value of φ on $\mathtt{obs}_0$ to be $\mathtt{1}$. Thus, determining the value of a formula on $\mathtt{obs}_0$ simply by counting the number of examples masked by $\mathtt{obs}_0$ that exhibit a certain property is not meaningful.

Following the argument above, a refined version of the objection could be raised. Consider a probability distribution $\mathcal{D}$ and a masking process mask such that

$$Pr\big[\mathtt{val}(\varphi \mid \mathtt{exm}) = \mathtt{1} \mid \mathtt{exm} \leftarrow \mathcal{D};\, \mathtt{obs} \leftarrow \mathtt{mask}(\mathtt{exm});\, \mathtt{obs} = \mathtt{obs}_0\big] \geqslant 0.999;$$

that is, in those cases that the particular observation $\mathtt{obs}_0$ is drawn, formula φ evaluates to $\mathtt{1}$ on the underlying example with overwhelming probability. This situation is often illustrated via a toy domain of observing birds, without, however, observing whether the birds are penguins. In this domain the underlying probability distribution $\mathcal{D}$ is taken to be such that most observed birds are not penguins, and thus have the ability to fly. The question, then, is whether it would be more reasonable to define the value of φ (the formula that we employ to make predictions on the ability of observed birds to fly) on the observation $\mathtt{obs}_0$ of the bird Tweety, to be $\mathtt{1}$. Our answer remains no. There exists no theoretical justification as to why a probability of 0.999, or any other probability for that matter, is high enough for a prediction of $\mathtt{1}$ to be made over a "don't know" prediction. It is easy to devise scenarios where it is preferable for an agent to predict "don't know" and be aware of this lack of certainty, rather than predicting a $\{\mathtt{0},\mathtt{1}\}$ value and risking a wrong prediction without knowing that this is happening. In such scenarios it is preferable for a "don't know" prediction to be made, for the lack of certainty to be recorded, and only then for the "don't know" prediction to be replaced with a $\{\mathtt{0},\mathtt{1}\}$ value by the agent's deliberation mechanism, in case such a value is believed to be likely true.

Suppose that we even subscribe to the view that for all practical purposes a probability of, say, at least 0.75 would be appropriate for φ to predict $\mathtt{1}$, and a probability of, say, at most 0.25 would be appropriate for φ to predict $\mathtt{0}$; φ would predict "don't know" only in the remaining cases. That is, suppose that for some domains it is more preferable to risk making a wrong prediction with a small probability, over making a "don't know" prediction. This setting is still not meaningful. An agent has no access to the underlying probability distribution $\mathcal{D}$ from which examples are drawn, and thus the problem of determining the probability in question given only $\mathtt{obs}_0$ is generally impossible. That is, given that $\mathtt{val}(\varphi \mid \mathtt{obs}_0) = *$, it is not possible to estimate what the risk of making a wrong prediction is, and thus there is no argument in favor of choosing to make a $\{\mathtt{0},\mathtt{1}\}$ prediction over the "don't know" prediction that is suggested by $\mathtt{val}(\varphi \mid \mathtt{obs}_0) = *$.

A second objection relates to our choice of how to measure the degree of accuracy. Recall that the degree of accuracy of a formula φ is the probability with which it predicts the correct underlying value of the target attribute, or it predicts "don't know". It could be argued that the degree of accuracy should be computed with respect only to the $\{\mathtt{0},\mathtt{1}\}$ predictions made by a formula, ignoring all "don't know" predictions; that is, the degree of accuracy should be defined to be the percentage of correct predictions *among* the $\{\mathtt{0},\mathtt{1}\}$ predictions. After all, would it not be more natural to account for the percentage of "don't know" predictions by introducing a second metric, call it the degree of completeness of a formula, that captures the probability with which $\{\mathtt{0},\mathtt{1}\}$ predictions are made? We agree that this alternative approach does have an appeal. However, we opted not to follow it for two main reasons:

(i) The degree of completeness $1-\omega$ of a formula, the degree of its accuracy $1-\varepsilon$ under our proposed definition, and the degree of its accuracy $1-\varepsilon'$ under the alternative definition discussed above, can be trivially derived from each other, since $\varepsilon' = \varepsilon/(1-\omega)$;

(ii) While the alternative degree of accuracy would not be meaningful by itself without the associated degree of completeness, our proposed degree of accuracy encompasses both metrics in one, building on the existence of a natural degree of completeness for a formula, as this follows from the fact that formulas cannot choose to abstain from making $\{\mathtt{0},\mathtt{1}\}$ predictions, and predict $*$ only when sufficient information is missing for such a "don't know" prediction to be justified (cf. the first objection).

A third objection relates to our choice of the requirements for learnability. Recall that we require the returned hypothesis to be highly accurate only. It could be argued that among the formulas that achieve the same degree of accuracy, one should prefer those formulas that have a higher degree of completeness (cf. the second objection), punishing, thus, the formulas that predict "don't know" more often. For instance, consider the case where the target concept is a parity φ that depends on a strict subset of all the attributes $\mathcal{A}$, and assume that the masking process is such that exactly one attribute is masked in each observation. Clearly, the parity φ_0 over all the attributes $\mathcal{A}$ will always predict "don't know", and will, thus, be 1-accurate. Note, however, that the parity φ is also 1-accurate, and does not always predict "don't know" since the formula φ might not depend on the attribute that is masked in some observations. Should we not prefer that a learning algorithm returns φ instead of φ_0? Yes, we should. However, we argue that the way to do this is not by imposing an additional requirement on the learner, but by restricting the hypothesis class to capture our prior knowledge that the target concept does not depend on all the attributes $\mathcal{A}$. This would exclude the parity φ_0 from being considered as a possible hypothesis. Looking at the same argument from a different angle, if some target concept cannot be a priori excluded (and be removed from the hypothesis class), then it is not possible during the learning phase to make a case for one hypothesis over another; after all, the hypothesis that makes more "don't know" predictions might be the actual target concept that the learner is looking for.

It could, nonetheless, be argued that a learning algorithm need not necessarily identify the actual target concept, but *any* hypothesis that is highly accurate. Among those that satisfy this requirement, then, would it not be meaningful to insist that the returned hypothesis is as complete as possible? We agree that it would be *desirable* to obtain a hypothesis that achieves the highest possible degree of completeness. At the same time, however, we note that *insisting* that this be the case in general is not reasonable. Unlike the feedback that observations provide on how the accuracy of a hypothesis compares to the optimal accuracy (which is achieved by the target concept), observations provide no indication on how the completeness of a hypothesis compares to the optimal completeness (which is achieved by some hypothesis that might differ from the target concept). In the general case, then, one cannot expect a learning algorithm to provide the same type of guarantees for completeness as it does for accuracy. The characterization of domains for which certain completeness guarantees on the learned hypothesis could be meaningfully insisted upon remains open.

3.2. Qualitative characteristics of masking

In addition to their concealment degree, masking processes may be characterized based on qualitative criteria. We briefly discuss next some representative types of masking processes along two dimensions that have been traditionally considered in the Statistical Analysis literature [16,26]. In the first dimension, the *pattern* of masked attributes is considered: in a *univariate* pattern only a single attribute is masked, or more generally, attributes in a fixed set are either all masked or all non-masked; in an *arbitrary* pattern any of the attributes may be masked without any constraints.[2] In the second dimension, the nature of the *dependence* of the masked attributes on the underlying examples is considered: attributes are masked *completely at random* (*MCAR*) if the masking of attributes is independent of the underlying example; attributes are masked *at random* (*MAR*) if the masking of attributes may depend on the values of the non-masked attributes only; finally, attributes are masked *not at random* (*MNAR*) if the masking of attributes depends on the values of the masked attributes in the underlying example.

In the simplest scenario (Type 1: univariate pattern and MCAR), a masking process mask maps examples to observations that mask only some target attribute x_t, and this happens randomly with some fixed probability p, and independently of the example being mapped. Since every accuracy conflict is observed with probability at least $1 - p$, and since there are accuracy conflicts that are observed with probability at most $1 - p$, it follows that mask is p-concealing for the target attribute x_t w.r.t. any class $\mathcal{F}$ of formulas.

The preceding scenario captures situations where some component in an agent's sensors randomly fails to provide a reading. In a natural variation (Type 2: univariate pattern and MAR (possibly not MCAR)), a component does not provide a reading in a manner possibly dependent on the readings of other components, but still independent of its own reading. The degree of concealment of masking processes with such dependence properties may take any value in the range $[0, 1]$.

In a different scenario (Type 3: arbitrary pattern and MCAR with independent masking), each attribute x_i is masked in observations randomly with some fixed probability p_i, which may differ across attributes. The attributes are still masked independently of the underlying example, and of each other. For any given target attribute x_t, formulas may now predict "don't know" on some observations, since the remaining attributes may also be masked. This fact seemingly makes the calculation of the concealment degree of a Type 3 masking process mask for x_t more involved than for a Type 1 masking process. Yet, the independence of masking across attributes, imposed by the product distribution, implies that mask is, in fact, p_t-concealing for each particular target attribute x_t w.r.t. any class $\mathcal{F}$ of formulas.

It is possible for attributes to be masked independently of the underlying example, but not from each other (Type 4: arbitrary pattern and MCAR with correlated masking). For instance, a masking process could map examples to observations so that with probability 0.23 the first half of the attributes are masked, with probability 0.36 those attributes indexed with a prime number are masked, and with probability 0.41 attribute x_{4397} is masked. Due to the correlation, the evaluation of the concealment degree of this masking process is far from straightforward, and depends on the target attribute x_t, and the class $\mathcal{F}$ of formulas.

In the most general case (Type 5: arbitrary pattern and MNAR), each attribute x_i is masked in observations in a manner that depends on its value in the underlying example. The human eyes exhibit such a behavior, by masking readings that are very bright, through the closing of the eyelids. A survey, viewed as a sensor, also behaves thus, since the lack of response in a question may be correlated with the answer. Masking processes of this type may hide information adversarially, and may have a high degree of concealment.

Although not an exhaustive list of types, the aforementioned discussion illustrates that the qualitative characteristics of a masking process are largely orthogonal to its degree of concealment, and to the ease with which the degree of concealment may be calculated. To emphasize this point further, consider the *deterministic* masking process mask that maps examples to observations so that some target attribute x_t is either always or never masked, depending on whether the Goldbach Conjecture is true or false under the standard axioms of mathematics. It is easy to see that attributes in observations are masked completely at random in a univariate pattern, yet, the concealment degree of mask for x_t w.r.t. any class $\mathcal{F}$ of formulas is either 0 or 1. In fact, as far as we know, the truth of the Goldbach Conjecture might be independent of the standard axioms of mathematics, in which case so would be the actual concealment degree of mask, meaning that it would be impossible to mathematically prove what the concealment degree of mask is.

3.3. Semantics of "don't know" predictions

Recall that in autodidactic learning, a formula makes a "don't know" prediction if and only if insufficient information exists in an observation for the formula to be evaluated — we have already defended this choice earlier in this section. Thus, although a learner may return a hypothesis that abstains from making predictions in certain observations, the abstention is beyond the control of the learner or the hypothesis, and depends only on the masking process and the observations to which it gives rise.

[2] A third possibility exists, that of a *monotone* pattern. This pattern appears in certain statistical studies where an attribute masked in some observation remains so in all subsequent observations. Monotone patterns are not meaningful in the setting we consider, where observations are assumed to be drawn independently from each other.

An immediate alternative is to consider a setting where the learner returns a program, rather than a hypothesis. The program receives as input an observation, and *decides* whether to predict "don't know", or a $\{0, 1\}$ value. In this setting, the program may decide to predict "don't know" even on observations that offer complete information on all attributes. It is then evident that simply measuring how accurate the predictions of a program are is not sufficient, as the program may simply choose to always abstain from making predictions. A second metric of completeness is needed, that measures how often "don't know" predictions are made. In a PAC-like model, then, one would expect not only that the degree of accuracy of the program is sufficiently high, but that the same is true also for the degree of completeness of the program. Overall, we would expect a learner to produce a program that makes only a few "don't know" predictions, and only a few inaccurate $\{0, 1\}$ predictions. Put differently, we may simply consider both "don't know" predictions and inaccurate predictions as wrong predictions, and then ask that the program produced by a learner makes only a few such wrong predictions. Rivest and Sloan [24] consider a treatment of when "don't know" predictions are made similar to that discussed above, in the context of complete information.

In a second alternative, the learner is expected to produce a hypothesis (or program) that never predicts "don't know". Thus, a $\{0, 1\}$ prediction is made even on observations that offer incomplete information for the value of the target attribute to be uniquely determined. The PAC-like guarantees that one may expect in such a setting is to achieve a probability of making an accurate prediction that is sufficiently close to the information-theoretically optimal probability that can be achieved. Since the value of the target attribute does not follow deterministically from the available information, the information-theoretically optimal probability is, in general, not 1. Such a treatment is followed, for instance, by Kearns and Schapire [13].

A third alternative exists also, where the goal of a learner is to produce a hypothesis that given an observation predicts whether the target attribute is masked in that observation. Thus, the hypothesis no longer attempts to accurately recover the missing value of the target attribute. The learning goal is no longer that of identifying the structure that exists in the underlying examples. Instead, the goal is that of identifying the structure that exists in the observations. In some sense, then, in this setting one assumes that the way information is sensed by an agent, including what this information is and what parts of it are missing, is structured. For instance, the target attribute might be masked if and only if certain other attributes are masked. Note that this setting resembles the case of learning from complete information as in the original PAC model, with the only difference that instead of learning a boolean formula over boolean attributes, one learns a ternary formula over ternary attributes. This approach has been taken, for instance, in the work of Valiant [29] on Robust Logics, and the work of Goldman et al. [9].

3.4. Degree of supervision while learning

The focus of the autodidactic learning model is to provide a framework for studying what can be learned in a truly autonomous manner, where no teacher is ever present. This, in particular, implies that the value of the target attribute might not be always available, not even during the learning phase. Varying the degree of availability of the value of the target attribute gives rise to various settings that relate to certain learning models in the literature. The orthogonal issue of how non-target attributes are masked is dealt with later on.

On the one extreme, one may consider a setting where the target attribute is always masked in the observations available to a learner. According to the autodidactic learning model, in any non-trivial setting the degree of concealment of the masking process that gives rise to these observations is 1. It is, thus, impossible to learn to predict the missing value of the target attribute accurately (cf. Theorem 2.1).

Such a scenario may be contrasted to the unsupervised learning model, where it is also the case that the target attribute is always masked. Despite this lack of information, in unsupervised learning one is expected to group the available observations in meaningful clusters that maximize some metric. The fundamental difference between the goal of unsupervised learning and autodidactic learning is that the former does not attempt to uncover some hidden, but definite, reality about the value of the target attribute. It simply partitions observations into clusters without associating each cluster with a value for the target attribute. That is, even if the clusters are identified perfectly, i.e., all observations where the masked target attribute has a hidden value of 0 are grouped together, it is impossible to know whether a given cluster corresponds to the target attribute having a 0 value or a 1 value. Overall, then, the focus of unsupervised learning is different from that of autodidactic learning, and the predictive guarantees that unsupervised learning offers are unrelated to those required for the task of information recovery that is examined in this work.

On the other extreme, one may consider a setting where the target attribute is never masked in the observations available to a learner. According to the autodidactic learning model, the degree of concealment of the masking process that gives rise to these observations is 0. Thus, learning to predict the missing value of the target attribute accurately is not compromised, and, in fact, accuracy degenerates to consistency.

The assumption that the value of the target attribute is available during the learning phase is that followed by supervised learning models. Similar to autodidactic learning, supervised learning seeks to identify how the value of the target attribute is determined by the values of the rest of the attributes. The fundamental difference between the goal of supervised learning and autodidactic learning lies in the availability of the value of the target attribute during the *evaluation* phase. In autodidactic learning, the observations during the evaluation phase come from the same source as the observations during the learning phase, namely the sensors of an agent. Thus, by assuming that the target attribute is never masked in observations during the learning phase, it follows that the same is true during the evaluation phase. This implication, then, makes the goal of learning to predict the value of the target attribute superfluous; the value of the target attribute is always observed, and need not be predicted. In supervised learning, on the other hand, it is valid to assume that the value of the target attribute is not available during the evaluation phase; thus, the goal of learning to predict the value of the target value is meaningful. The natural way to interpret the discrepancy between the learning and the evaluation phase in supervised learning is to take the approach that the learner's sensors never provide the agent with the value of the target attribute. However, during the learning phase an external teacher supervises the learner and provides it with the values of the target attribute. Overall, then, the premises of supervised learning are in contrast to our goal of developing a framework for learning in a completely autonomous manner. Yet, this does not restrict us from using within the context of autodidactic learning, techniques and algorithms that were developed for supervised learning models.

3.5. Regularity on how information is hidden

In a truly autonomous learning setting, where no teacher is available to offer the agent sufficient information to aid the learning task, few or no assumptions can be made as to how information is hidden in observations when an agent senses its environment. In autodidactic learning this is reflected by making no assumptions on what the masking process looks like; in this respect, we follow the treatment of the PAC model, where no assumption is made on what the distribution over examples is, and thus, on how nature chooses the states of the environment that will be sensed by the agent. Nonetheless, in certain domains, most notably those involving a teacher, some regularity may be assumed on the sensors of an agent; more precisely, the regularity might be a result of the combined workings of an agent's sensors and a teacher that completes some of the information missing in the agent's sensory inputs (cf. the discussion on supervised learning earlier in this section). In these teacher-based settings, thus, the regularity of how information is hidden often differs between the learning and the evaluation phase. We examine next the assumptions that some learning models place on the regularity of how information is hidden *during the learning phase*.

On the one extreme we have learning models where complete observations are assumed. The original PAC model is often taken to fall in this category, as well as most of its variants in the Learning Theory literature (e.g., [24]). Despite this, Valiant's seminal paper that introduced the PAC model [28] also discusses learnability from partial observations, where a certain benign type of missing information in observations is considered. It is assumed that information may be hidden on non-target attributes as long as the non-masked attributes provide enough information for the target concept to evaluate to `1` when the value of the target attribute is `1` in an observation. So, examples in which the value of the target attribute is `1` are masked by observations in a manner that no *essential* information is lost, whereas examples in which the value of the target attribute is `0` might be masked by arbitrary observations.

We are not aware of any other learning models beyond autodidactic learning that lie on the other end of the spectrum, and make essentially no assumptions on how information is missing. Among the approaches that we are aware of, conceptually closer to autodidactic learning are learning models that assume some independence on the way the various attributes are masked. Such is the case in the work of Decatur and Gennaro [6], where each attribute is masked in observations randomly and independently of the underlying example and other attributes, with some fixed probability p that is constant across attributes.

Other learning models make no assumptions on how information is hidden in observations *for non-target attributes*, but restrict the target attribute is some way that depends on the masked non-target attributes. This is the case in the model of learning from examples with unspecified attribute values [9], where the target attribute is masked if and only if the value of the target concept cannot be determined by the values that are available on the rest of the attributes. This asymmetry is best explained by the presence of a teacher that ensures that this constraint is met. The Robust Logics framework [29] does not a priori impose a similar restriction on the masking of any particular target attribute, but it implicitly assumes that this is the case for learnability to be possible. In both cases, a learner attempts to learn to predict when the value of the target attribute is "don't know" by learning the structure of *observations*, and by allowing hypotheses to condition their predictions on whether the values of certain non-target attributes are "don't know". Contrast this to the autodidactic learning model where the goal is to learn the structure of the *underlying examples*, and where the "don't know" values of attributes cannot be explicitly taken into account by hypotheses.

Schuurmans and Greiner [27] consider a model where the target attribute is never masked, and examine various cases on how the remaining attributes are masked: arbitrarily, according to some product distribution (cf. Type 3 masking in Section 3.2), or not masked at all. Ben-David and Dichterman [2] consider a different setting where k attributes are not masked in each observation, but the choice of which attributes are not masked may change across observations, and can be actively chosen by the learner.

A number of other models that on the surface consider complete observations, can be effectively viewed as dealing with a special class of partial observations, where the masked attributes are those in a fixed unknown subset. If one attempts then to compute the value of the target attribute as a function of the values of the *non-masked attributes*, then one can interpret the lack of existence of a deterministic function as being due to either classification noise [1], the existence of only a probabilistic function [13], or the existence of a deterministic function that is occasionally switched [3]. Fig. 1 shows how missing information may manifest itself in these three ways.

MANIFESTATIONS OF MISSING INFORMATION Assume that the set of attributes required to fully describe an environment is partitioned into a set of hidden attributes $\mathcal{A}' = \{x'_1, x'_2, \ldots, x'_{|\mathcal{A}'|}\}$, always assigned a "don't know" value in observations, and a set of known attributes $\mathcal{A} = \{x_1, x_2, \ldots, x_{|\mathcal{A}|}\}$, always assigned a $\{0, 1\}$ value in observations. Let $\texttt{obs} \in \{0, 1\}^{|\mathcal{A}|}$ be a complete observation, $\texttt{obs}' \in \{0, 1, *\}^{|\mathcal{A}' \cup \mathcal{A}|}$ be the unique corresponding partial observation, and $\texttt{exm}' \in \{0, 1\}^{|\mathcal{A}' \cup \mathcal{A}|}$ be an example among those masked by $\texttt{obs}'$ that is randomly drawn from some underlying probability distribution. Then:

(i) When the target attribute $x_t \in \mathcal{A}$ is expressed by the target concept $\varphi' \leftrightarrow \varphi$, where φ' is a formula over $\mathcal{A}'$, and φ is a formula over $\mathcal{A}$, then partial observations over $\mathcal{A}' \cup \mathcal{A}$ may manifest themselves as complete observations over $\mathcal{A}$, with x_t being noisily expressed by the manifested target concept φ. The hidden value of φ' on $\texttt{exm}'$ determines whether x_t obtains the value of the manifested target concept φ, or a noisy value, on the manifested complete observation $\texttt{obs}$.

(ii) When the target attribute $x_t \in \mathcal{A}$ is expressed by the target concept φ', where φ' is a formula over $\mathcal{A}'$, then partial observations over $\mathcal{A}' \cup \mathcal{A}$ may manifest themselves as complete observations over $\mathcal{A}$, with x_t being expressed by a manifested probabilistic target concept that evaluates to $\texttt{1}$ on $\texttt{obs}$ with some probability $p_{\varphi'}(\texttt{obs})$. The probability with which the hidden value of φ' is $\texttt{1}$ on $\texttt{exm}'$ determines the probability $p_{\varphi'}(\texttt{obs})$ with which x_t obtains the value $\texttt{1}$ on the manifested complete observation $\texttt{obs}$.

(iii) When the target attribute $x_t \in \mathcal{A}$ is expressed by the target concept $\bigwedge_{j=1}^{k}(\varphi'_j \to \varphi_j)$, where $\{\varphi'_1, \varphi'_2, \ldots, \varphi'_k\}$ is a set of formulas over $\mathcal{A}'$ exactly one of which evaluates to $\texttt{1}$ on any given truth-assignment to the attributes $\mathcal{A}'$, and $\{\varphi_1, \varphi_2, \ldots, \varphi_k\}$ is a set of formulas over $\mathcal{A}$, then partial observations over $\mathcal{A}' \cup \mathcal{A}$ may manifest themselves as complete observations over $\mathcal{A}$, with x_t switching between being expressed by one of the manifested target concepts in $\{\varphi_1, \varphi_2, \ldots, \varphi_k\}$. The unique formula φ'_j whose hidden value is $\texttt{1}$ on $\texttt{exm}'$ determines the manifested target concept φ_j of which x_t obtains the value on the manifested complete observation $\texttt{obs}$.

Fig. 1. Various ways in which missing information may manifest itself.

4. Learnability results and tools

We continue in this section to establish some learnability results in the autodidactic learning model. In the spirit of the PAC model, which the autodidactic learning model extends, the goal is to establish that certain concept classes are learnable, despite the arbitrary manner in which information is hidden in observations. Since our ultimate goal is that of predicting accurately the value of the target attribute, we expect learned hypotheses to be highly accurate. By Theorem 2.2, then, it suffices to learn highly consistent hypotheses.

Our learnability results are derived through existing PAC learning algorithms and techniques, which we extend to the case of partial observability. As a motivating example of how this can be done, we consider the following algorithm for properly learning monotone conjunctions under the PAC semantics.

> Consider only those observations that assign the value $\texttt{1}$ to the target attribute x_t. Out of the attributes in $\mathcal{A} \setminus \{x_t\}$, remove those that are assigned the value $\texttt{0}$ in any observation during the learning phase. Return the hypothesis comprised of the conjunction of all remaining attributes.

This algorithm was proposed and was proved correct under the PAC semantics in Valiant's seminal paper [28]. The idea behind the algorithm is essentially the following: identify a hypothesis that agrees with all given learning examples, and then appeal to an Occam's Razor type of argument.

Consider the application of the algorithm above on partial observations. Clearly, observations $\texttt{obs}$ with $\texttt{obs}[t] = *$ are ignored. Since this is also the case for observations $\texttt{obs}$ with $\texttt{obs}[t] = \texttt{0}$, one could safely replace all $*$ values of the target attribute x_t in observations, with the value $\texttt{0}$, without affecting the outcome of the algorithm. By an entirely analogous argument, replacing all $*$ values of the attributes $\mathcal{A} \setminus \{x_t\}$ in observations, with the value $\texttt{1}$, would not affect the outcome of the algorithm, since those attributes are ignored by the algorithm. Overall, there exist certain "default" values that may be assigned to masked attributes so that the algorithm will face *complete* observations, *without* this affecting the algorithm's behavior. This suggests that it may be possible to obtain consistent learners by reducing the learning problem to one over complete observations. It suggests also that the principle known as Occam's Razor may apply to the case of partial observations. Consequently, certain positive results and certain learnability techniques under the PAC semantics may be lifted to the case of consistent learnability from partial observations.

4.1. Learnability through Occam's Razor

Intuitively, one may see that the ideas behind Occam's Razor [4] do not rely on the learning observations being complete. For completeness of the presentation, we reproduce below one version of Occam's Razor for the case of consistent learnability. Although we do not actually employ this technique to derive our positive learnability results later on, some of those results could have also been derived through an Occam's Razor argument. It remains an interesting prospect to derive novel autodidactic learnability results that cannot be established through the other techniques that we consider in this section.

Definition 4.1 *(Compressibility).* An algorithm $\mathcal{L}$ is a ***compressor for*** a learning task $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ over $\mathcal{A}$ if there exists a real number $\beta \in [0,1)$ such that for every sample $\mathcal{O}$ of observations given which a formula $c \in \mathcal{C}$ is 1-consistent with x_t, every real number $\delta \in (0,1]$, and every real number $\varepsilon \in (0,1]$, algorithm $\mathcal{L}$ has the following property: given access to $\mathcal{A}$, $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$, δ, ε, and $\mathcal{O}$, algorithm $\mathcal{L}$ runs in time polynomial in $1/\delta$, $1/\varepsilon$, $|\mathcal{A}|$, *size*(c), and $|\mathcal{O}|$, and returns, with probability $1-\delta$, a hypothesis $h \in \mathcal{H}$ that is $(1-\varepsilon)$-consistent with x_t given $\mathcal{O}$, and its size is linear in $|\mathcal{O}|^\beta$ and polynomial in $1/\delta$, $1/\varepsilon$, $|\mathcal{A}|$, and *size*(c). The concept class $\mathcal{C}$ over $\mathcal{A} \setminus \{x_t\}$ is ***compressible on*** the target attribute $x_t \in \mathcal{A}$ ***by*** the hypothesis class $\mathcal{H}$ over $\mathcal{A} \setminus \{x_t\}$ if there exists a compressor for $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ over $\mathcal{A}$.

Unlike a consistent learner (cf. Definition 2.3), a compressor is not given oracle access to observations, but instead a sample $\mathcal{O}$ of observations, with no mention of an underlying reality from which observations are obtained. The consistency guarantees of the returned hypothesis are expected to be with respect to the given sample, while it is assumed that a perfectly consistent formula from the concept class exists. The compressor is also allowed to expend resources that increase with the size of $\mathcal{O}$. The compression requirement is accounted for by insisting that the size of the returned hypothesis grows only linearly in $|\mathcal{O}|^\beta$, for some non-negative constant β less than 1. We next establish that a compressor for a learning task is essentially a consistent learner for that learning task.

Theorem 4.1 *(Learning through compression). Consider a learning task $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ over $\mathcal{A}$. The concept class $\mathcal{C}$ is consistently learnable on x_t by $\mathcal{H}$ if the concept class $\mathcal{C}$ is compressible on x_t by $\mathcal{H}$.*

Proof. Consider an algorithm $\mathcal{L}'$ that is a compressor for the learning task $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ over $\mathcal{A}$. We construct an algorithm $\mathcal{L}$ as follows. Fix a probability distribution $\mathcal{D}$ supporting $\mathcal{C}$ for x_t, with c being the target concept for x_t under $\mathcal{D}$, a masking process `mask`, a real number $\delta \in (0,1]$, and a real number $\varepsilon \in (0,1]$. Then, algorithm $\mathcal{L}$, given access to $\mathcal{A}$, $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$, δ, ε, and an oracle returning observations drawn from `mask`$(\mathcal{D})$, proceeds as follows:

> Algorithm $\mathcal{L}$ draws a sample $\mathcal{O}$ of a number of observations (to be determined later) from the oracle, and simulates algorithm $\mathcal{L}'$ with input $\mathcal{A}$, $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$, $\delta' = \delta/2$, $\varepsilon' = \varepsilon/2$, and $\mathcal{O}$. When algorithm $\mathcal{L}'$ returns a hypothesis h, algorithm $\mathcal{L}$ returns the hypothesis h, and terminates.

We now prove that algorithm $\mathcal{L}$ is a consistent learner for the learning task $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ over $\mathcal{A}$. To do so, we prove that, with probability $1-\delta$, the returned hypothesis h is $(1-\varepsilon)$-consistent with x_t under $\mathcal{D}$ and `mask`, and that algorithm $\mathcal{L}$ runs in time polynomial in $1/\delta$, $1/\varepsilon$, $|\mathcal{A}|$, and *size*(c).

By construction of algorithm $\mathcal{L}$, the simulated algorithm $\mathcal{L}'$ is given access to the sample $\mathcal{O}$. Clearly, $c \in \mathcal{C}$ is 1-consistent with x_t given $\mathcal{O}$. By the choice of δ' and ε', and by Definition 4.1, there exists a constant $\beta \in [0,1)$ such that algorithm $\mathcal{L}'$ runs in time polynomial in $2/\delta$, $2/\varepsilon$, $|\mathcal{A}|$, *size*(c), and $|\mathcal{O}|$, and returns, with probability $1-\delta/2$, a hypothesis $h \in \mathcal{H}$ that is $(1-\varepsilon/2)$-consistent with x_t given $\mathcal{O}$, and its size is linear in $|\mathcal{O}|^\beta$ and polynomial in $2/\delta$, $2/\varepsilon$, $|\mathcal{A}|$, and *size*(c).

Consider the set $\mathcal{H}'$ of all formulas in $\mathcal{H}$ that are $(1-\varepsilon/2)$-consistent with x_t given $\mathcal{O}$, and their size is linear in $|\mathcal{O}|^\beta$ and polynomial in $2/\delta$, $2/\varepsilon$, $|\mathcal{A}|$, and *size*(c). By Theorem 2.3, and setting $\gamma = \varepsilon/2$, each formula $\varphi \in \mathcal{H}'$ is, except with probability $e^{-|\mathcal{O}|\varepsilon^2/2}$, such that φ is $(1-\varepsilon)$-consistent with x_t under $\mathcal{D}$ and `mask`. By a union bound, except with probability $|\mathcal{H}'|e^{-|\mathcal{O}|\varepsilon^2/2}$, every formula $\varphi \in \mathcal{H}'$ is $(1-\varepsilon)$-consistent with x_t under $\mathcal{D}$ and `mask`. Overall, algorithm $\mathcal{L}$ will return, except with probability $\delta/2 + |\mathcal{H}'|e^{-|\mathcal{O}|\varepsilon^2/2}$, a hypothesis $h \in \mathcal{H}$ that is $(1-\varepsilon)$-consistent with x_t under $\mathcal{D}$ and `mask`.

Since $\mathcal{H}'$ contains formulas with size only linear in $|\mathcal{O}|^\beta$ and polynomial in $2/\delta$, $2/\varepsilon$, $|\mathcal{A}|$, and *size*(c), it follows that $\log|\mathcal{H}'| \leqslant m|\mathcal{O}|^\beta \cdot \mathrm{poly}(2/\delta, 2/\varepsilon, |\mathcal{A}|, \mathit{size}(c))$, for some constant m. Thus, the probability that algorithm $\mathcal{L}$ will return a hypothesis that is not $(1-\varepsilon)$-consistent with x_t under $\mathcal{D}$ and `mask`, is at most

$$\delta/2 + 2^{m|\mathcal{O}|^\beta \cdot \mathrm{poly}(2/\delta, 2/\varepsilon, |\mathcal{A}|, \mathit{size}(c))} e^{-|\mathcal{O}|\varepsilon^2/2} \leqslant \delta/2 + 2^{-|\mathcal{O}|^\beta (|\mathcal{O}|^{1-\beta}\varepsilon^2/2 - m\cdot \mathrm{poly}(2/\delta, 2/\varepsilon, |\mathcal{A}|, \mathit{size}(c)))}.$$

Fixing $|\mathcal{O}|$ to be the least positive integer that exceeds the quantity

$$\left(\frac{2}{\varepsilon^2}\left(\log\frac{2}{\delta} + m \cdot \mathrm{poly}\big(2/\delta, 2/\varepsilon, |\mathcal{A}|, \mathit{size}(c)\big)\right)\right)^{\frac{1}{1-\beta}},$$

trivially implies that $|\mathcal{O}|^\beta \geqslant 1$, and also ensures that $|\mathcal{O}|^{1-\beta}\varepsilon^2/2 - m \cdot \mathrm{poly}(1/\delta, 1/\varepsilon, |\mathcal{A}|, \mathit{size}(c)) \geqslant \log\frac{2}{\delta}$, making the probability that algorithm $\mathcal{L}$ will return a hypothesis that is not $(1-\varepsilon)$-consistent with x_t under $\mathcal{D}$ and `mask`, be at most δ, as required.

The running time of algorithm $\mathcal{L}$ comprises the time required to draw the sample $\mathcal{O}$ of observations, and the time required to simulate algorithm $\mathcal{L}'$. Both tasks are carried out in time polynomial in $2/\delta$, $2/\varepsilon$, $|\mathcal{A}|$, *size*(c), and $|\mathcal{O}|$, and since $|\mathcal{O}|$ is also polynomial in $1/\delta$, $1/\varepsilon$, $|\mathcal{A}|$, and *size*(c), the claim follows. □

4.2. Learnability through reductions

Efficiently reducing one problem to another is a natural approach to establish that solving the first problem is not harder than solving the second one. Thinking of each input of a problem as corresponding to an instance of that problem, we term the mapping between inputs an ***instance mapping***. Thinking of an output of a problem as corresponding to a solution of an instance of that problem, we term the mapping between outputs a ***solution mapping***. Such reductions have been widely employed within the area of computational complexity, but also within the context of PAC learning [23], where the oracle available to the first learner is transformed to an oracle to be made available to the second learner, while the returned hypotheses of the second learner are transformed to hypotheses to be returned by the first learner. Other inputs required during learning (e.g., δ, ε) may also be transformed through the instance mapping. In the case of reductions between learning problems, a solution mapping is referred to as a ***hypothesis mapping***.

The transformation of oracles in reductions in the context of PAC learning is relatively straightforward. Each complete observation drawn from the original oracle is mapped to another complete observation, and the latter one is thought of as being drawn from the transformed oracle. In the case of autodidactic learning, however, an oracle returns observations drawn from some probability distribution $\mathtt{mask}_1(\mathcal{D}_1)$, where $\mathtt{mask}_1$ is not necessarily the identity mapping. Transforming this oracle to another one could, in principle, be done by simply mapping each drawn observation to another observation. This, however, does not suffice. The induced probability distribution over the resulting observations needs to be expressible in the form $\mathtt{mask}_2(\mathcal{D}_2)$, so that the resulting observations may be thought of as masking examples drawn from some underlying distribution $\mathcal{D}_2$. The following definition captures this requirement. For generality, we define one-to-many reductions, where one learning problem may be transformed to many others. This generality is not invoked later on to obtain our positive learnability results. In fact, we are unaware of any learnability results that were obtained through a one-to-many reduction. Whether one-to-many reductions are more powerful than one-to-one reductions in the context of learnability remains an interesting open problem.

Definition 4.2 *(Reductions between learning tasks).* A learning task $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ over $\mathcal{A}$ is ***reducible to*** a set $\{\langle x_t^j, \mathcal{C}^j, \mathcal{H}^j\rangle$ over $\mathcal{A}^j\}_{j=1}^r$ of learning tasks if there exists a hypothesis mapping $g : \mathcal{H}^1 \times \cdots \times \mathcal{H}^r \to \mathcal{H}$, and for every j: $1 \leqslant j \leqslant r$ there exists an instance mapping $f^j : \{0,1,*\}^{|\mathcal{A}|} \to \{0,1,*\}^{|\mathcal{A}^j|}$, and for every probability distribution $\mathcal{D}$ supporting $\mathcal{C}$ for x_t, with c being the target concept for x_t under $\mathcal{D}$, and every masking process $\mathtt{mask}$, there exists a probability distribution $\mathcal{D}^j$ supporting $\mathcal{C}^j$ for x_t^j, with c^j being the target concept for x_t^j under $\mathcal{D}^j$, and a masking process $\mathtt{mask}^j$, such that:

(i) for every tuple $\langle h^1, \ldots, h^r\rangle \in \mathcal{H}^1 \times \cdots \times \mathcal{H}^r$, and every observation $\mathtt{obs} \in \{0,1,*\}^{|\mathcal{A}|}$, it holds that $g(\langle h^1, \ldots, h^r\rangle)$ has a consistency conflict with x_t w.r.t. $\mathtt{obs}$ only if there exists j: $1 \leqslant j \leqslant r$ such that h^j has a consistency conflict with x_t^j w.r.t. $f^j(\mathtt{obs})$;
(ii) for every j: $1 \leqslant j \leqslant r$, the induced probability distribution $f^j(\mathtt{mask}(\mathcal{D}))$ is equal to $\mathtt{mask}^j(\mathcal{D}^j)$;
(iii) each of the instance and hypothesis mappings is computable in time polynomial in $|\mathcal{A}|$, *size*(c), and the size of its input; both r, and *size*(c^j) for every j: $1 \leqslant j \leqslant r$, are polynomial in $|\mathcal{A}|$ and *size*(c).

Roughly speaking, the three conditions of Definition 4.2 correspond, respectively, to the following requirements: Condition (i) ensures that the transformations of inputs and outputs between the involved learning problems are such that highly consistent hypotheses in the resulting problems correspond to a highly consistent hypothesis in the original problem. At the same time, Condition (ii) ensures that the instance mappings respect the requirement that the resulting observations mask examples drawn from some appropriate probability distribution. Finally, Condition (iii) ensures that the entire reduction is carried out efficiently. One may note that Definition 4.2 does not dictate how parameters δ and ε, which are, also, part of the inputs of a learning problem, are transformed. Indeed, given that the three aforementioned conditions hold, δ and ε may always be transformed appropriately; the proof of the following result illustrates this.

Theorem 4.2 *(Learning through reductions). Consider a learning task $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ over $\mathcal{A}$ that is reducible to the set of learning tasks $\{\langle x_t^j, \mathcal{C}^j, \mathcal{H}^j\rangle$ over $\mathcal{A}^j\}_{j=1}^r$. The concept class $\mathcal{C}$ is consistently learnable on x_t by $\mathcal{H}$ if for every j: $1 \leqslant j \leqslant r$, the concept class $\mathcal{C}^j$ is consistently learnable on x_t^j by $\mathcal{H}^j$.*

Proof. Assume that for every j: $1 \leqslant j \leqslant r$ the concept class $\mathcal{C}^j$ is consistently learnable on x_t^j by $\mathcal{H}^j$, and let algorithm $\mathcal{L}^j$ be a learner for the learning task $\langle x_t^j, \mathcal{C}^j, \mathcal{H}^j\rangle$ over $\mathcal{A}^j$. Let $g : \mathcal{H}^1 \times \cdots \times \mathcal{H}^r \to \mathcal{H}$ be the hypothesis mapping, and for every j: $1 \leqslant j \leqslant r$, let $f^j : \{0,1,*\}^{|\mathcal{A}|} \to \{0,1,*\}^{|\mathcal{A}^j|}$ be the instance mapping, whose existence is guaranteed by Definition 4.2. We construct an algorithm $\mathcal{L}$ as follows. Fix a probability distribution $\mathcal{D}$ supporting $\mathcal{C}$ for x_t, with c being the target concept for x_t under $\mathcal{D}$, a masking process $\mathtt{mask}$, a real number $\delta \in (0,1]$, and a real number $\varepsilon \in (0,1]$. Then, algorithm $\mathcal{L}$, given access to $\mathcal{A}$, $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$, δ, ε, and an oracle returning observations drawn from $\mathtt{mask}(\mathcal{D})$, proceeds as follows:

> For every j: $1 \leqslant j \leqslant r$, algorithm $\mathcal{L}$ simulates algorithm $\mathcal{L}^j$ with input $\mathcal{A}^j$, $\langle x_t^j, \mathcal{C}^j, \mathcal{H}^j\rangle$, $\delta^j = \delta/r$, $\varepsilon^j = \varepsilon/r$, and an oracle returning observations. Whenever algorithm $\mathcal{L}^j$ accesses the oracle and requests an observation, algorithm $\mathcal{L}$ draws an observation $\mathtt{obs}$ from its own oracle, and passes $f^j(\mathtt{obs})$ to algorithm $\mathcal{L}^j$. When each simulated algorithm $\mathcal{L}^j$ returns a hypothesis h^j, algorithm $\mathcal{L}$ computes and returns the hypothesis $g(\langle h^1, \ldots, h^r\rangle)$, and terminates.

We now prove that algorithm $\mathcal{L}$ is a consistent learner for the learning task $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ over $\mathcal{A}$. To do so, we prove that, with probability $1-\delta$, the returned hypothesis $g(\langle h^1, \ldots, h^r\rangle)$ is $(1-\varepsilon)$-consistent with x_t under $\mathcal{D}$ and $\texttt{mask}$, and that algorithm $\mathcal{L}$ runs in time polynomial in $1/\delta$, $1/\varepsilon$, $|\mathcal{A}|$, and *size*(c).

By construction of algorithm $\mathcal{L}$, each simulated algorithm $\mathcal{L}^j$ is given access to the oracle $f^j(\texttt{mask}(\mathcal{D}))$. By Condition (ii) of Definition 4.2, there exists a probability distribution $\mathcal{D}^j$ supporting $\mathcal{C}^j$ for x_t^j, with c^j being the target concept for x_t^j under $\mathcal{D}^j$, and a masking process $\texttt{mask}^j$, such that $f^j(\texttt{mask}(\mathcal{D})) = \texttt{mask}^j(\mathcal{D}^j)$. By Definition 2.3, algorithm $\mathcal{L}^j$ runs in time polynomial in $1/\delta^j$, $1/\varepsilon^j$, $|\mathcal{A}^j|$, and *size*(c^j), and returns, with probability $1-\delta^j$, a hypothesis $h^j \in \mathcal{H}^j$ that is $(1-\varepsilon^j)$-consistent with x_t^j under $\mathcal{D}^j$ and $\texttt{mask}^j$. By a union bound it follows that with probability $1-\sum_{j=1}^r \delta^j = 1-\delta$, every algorithm $\mathcal{L}^j$ will return a hypothesis $h^j \in \mathcal{H}^j$ that is $(1-\varepsilon^j)$-consistent with x_t^j under $\mathcal{D}^j$ and $\texttt{mask}^j$. Assume, now, that $g(\langle h^1, \ldots, h^r\rangle)$ is not $(1-\varepsilon)$-consistent with x_t under $\mathcal{D}$ and $\texttt{mask}$. Thus, the probability that $g(\langle h^1, \ldots, h^r\rangle)$ has a consistency conflict with x_t w.r.t. an observation $\texttt{obs} \leftarrow \texttt{mask}(\mathcal{D})$ is more than ε. By Condition (i) of Definition 4.2, it follows that the probability that there exists j: $1 \leqslant j \leqslant r$ such that h^j has a consistency conflict with x_t^j w.r.t. $f^j(\texttt{mask}(\mathcal{D})) = \texttt{mask}^j(\mathcal{D}^j)$ is more than ε. By the pigeonhole principle it follows that there exists j: $1 \leqslant j \leqslant r$ such that the probability that h^j has a consistency conflict with x_t^j w.r.t. $f^j(\texttt{mask}(\mathcal{D})) = \texttt{mask}^j(\mathcal{D}^j)$ is more than $\varepsilon/r = \varepsilon^j$; so, h^j is not $(1-\varepsilon^j)$-consistent with x_t^j under $\mathcal{D}^j$ and $\texttt{mask}^j$. This event, however, happens with probability at most δ. Therefore, with probability $1-\delta$, the returned hypothesis $g(\langle h^1, \ldots, h^r\rangle)$ is $(1-\varepsilon)$-consistent with x_t under $\mathcal{D}$ and $\texttt{mask}$, which establishes the first claim.

The running time of algorithm $\mathcal{L}$ comprises the running time of the r simulated algorithms, the time required to simulate all the oracle calls of those algorithms through the application of the instance mappings, and the time required to obtain the hypothesis $g(\langle h^1, \ldots, h^r\rangle)$ through the application of the hypothesis mapping. By Condition (iii) of Definition 4.2, each of the instance mappings is computable in time polynomial in $|\mathcal{A}|$ and *size*(c); thus, the size $|\mathcal{A}^j|$ of the set of attributes in each of the resulting learning tasks is polynomial in $|\mathcal{A}|$ and *size*(c). By the same condition, the size *size*(c^j) of the target concept in each of the resulting learning tasks is also polynomial in $|\mathcal{A}|$ and *size*(c). Since the same condition implies that r is polynomial in $|\mathcal{A}|$ and *size*(c), it follows that both $1/\delta^j = r/\delta$ and $1/\varepsilon^j = r/\varepsilon$ are polynomial in $1/\delta$, $1/\varepsilon$, $|\mathcal{A}|$, and *size*(c). Therefore, the input of each simulated algorithm $\mathcal{L}^j$ is polynomial in $1/\delta$, $1/\varepsilon$, $|\mathcal{A}|$, and *size*(c), and since the running time of algorithm $\mathcal{L}^j$ is polynomial in its input, it is also polynomial in $1/\delta$, $1/\varepsilon$, $|\mathcal{A}|$, and *size*(c). This, in turn, implies that each algorithm $\mathcal{L}^j$ accesses its oracle a number of times that is polynomial $1/\delta$, $1/\varepsilon$, $|\mathcal{A}|$, and *size*(c); hence, all applications of the instance mapping f^j are computable in time polynomial $1/\delta$, $1/\varepsilon$, $|\mathcal{A}|$, and *size*(c). Furthermore, the running time of the simulated algorithms implies that the size of $\langle h^1, \ldots, h^r\rangle$ is polynomial in $1/\delta$, $1/\varepsilon$, $|\mathcal{A}|$, and *size*(c), and by Condition (iii) of Definition 4.2 the hypothesis mapping is computable in time polynomial in $1/\delta$, $1/\varepsilon$, $|\mathcal{A}|$, and *size*(c). In conclusion, algorithm $\mathcal{L}$ runs in time polynomial in $1/\delta$, $1/\varepsilon$, $|\mathcal{A}|$, and *size*(c). This concludes the proof. □

Our motivating example of an algorithm for learning monotone conjunctions suggests the special case of reductions where observations in all the resulting learning tasks are complete. In terms of Definition 4.2, this corresponds to having, for each resulting learning task, an instance mapping whose codomain is that of complete observations.

Definition 4.3 *(Total reductions between learning tasks).* As a special case of Definition 4.2, a reduction from a learning task $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ over $\mathcal{A}$ to a set of learning tasks $\{\langle x_t^j, \mathcal{C}^j, \mathcal{H}^j\rangle$ over $\mathcal{A}^j\}_{j=1}^r$ is ***total*** if for every j: $1 \leqslant j \leqslant r$, the instance mapping is of the form $f^j : \{0,1,*\}^{|\mathcal{A}|} \to \{0,1\}^{|\mathcal{A}^j|}$.

Establishing total reductions is of particular interest for two reasons: (i) from a philosophical point of view, total reductions establish links between partial and complete observability, allowing one to identify conditions under which the lack of complete information does not affect learnability; (ii) from a more pragmatic point of view, these established links also relate autodidactic learnability to the well-studied PAC semantics, allowing one to carry positive results from the latter model to the former one.

4.3. Monotonicity preserves learnability

By using reductions we now establish a general result, showing that monotonicity of the concept class compensates for missing information in observations, in the sense that if a concept class of monotone formulas is learnable under the standard PAC semantics, then it remains so under the autodidactic learning semantics.

A formula φ over a set of attributes $\mathcal{A}$ is *monotone* if for every pair of examples $\texttt{exm}_1, \texttt{exm}_2 \in \{0,1\}^{|\mathcal{A}|}$ such that $\{i \mid x_i \in \mathcal{A}; \texttt{exm}_1[i] = 1\} \subseteq \{i \mid x_i \in \mathcal{A}; \texttt{exm}_2[i] = 1\}$, it holds that $\texttt{val}(\varphi \mid \texttt{exm}_1) \preceq \texttt{val}(\varphi \mid \texttt{exm}_2)$, where $\preceq$ imposes the natural ordering over the values $\{0,1\}$. In words, changing the input of a monotone formula so that more attributes are assigned the value 1 may result in the formula's value only remaining the same, or changing from 0 to 1. Consider, now, the value of a monotone formula φ on an observation $\texttt{obs}$. If $\texttt{val}(\varphi \mid \texttt{obs}) \in \{0,1\}$, then clearly mapping all attributes masked in $\texttt{obs}$ to any $\{0,1\}$ value will not affect the value of the formula. Furthermore, if $\texttt{val}(\varphi \mid \texttt{obs}) = *$, then mapping attributes masked in $\texttt{obs}$ either all to 0, or all to 1 will result in the formula obtaining the respective value. These simple properties suggest that partial observations may be replaced with complete observations so that the value of a monotone formula is affected in a predictable manner. This predictability, then, facilitates the existence of total reductions.

Theorem 4.3 *(Total self-reduction of monotone formulas). A learning task $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ over $\mathcal{A}$ is total reducible to a learning task $\langle x'_t, \mathcal{C}', \mathcal{H}'\rangle$ over $\mathcal{A}'$ such that $\mathcal{A}' = \mathcal{A}$, $x'_t = x_t$, $\mathcal{C}' = \mathcal{C}$, $\mathcal{H}' = \mathcal{H}$, and the hypothesis mapping is restricted to be the identity mapping, if the concept class $\mathcal{C}$ and the hypothesis class $\mathcal{H}$ are classes of monotone formulas, and $\mathcal{C}$ does not contain the tautology formula $\top$.*

Proof. We first define the constructs whose existence is required by Definition 4.2. Define the hypothesis mapping $g : \mathcal{H}' \to \mathcal{H}$ to be the identity mapping. Define the instance mapping $f : \{0, 1, *\}^{|\mathcal{A}|} \to \{0, 1\}^{|\mathcal{A}'|}$ so that for every observation $\mathtt{obs} \in \{0, 1, *\}^{|\mathcal{A}|}$, and every attribute $x'_i \in \mathcal{A}'$, it holds that: $f(\mathtt{obs})[i] = 0$ if $\mathtt{obs}[t] = *$; $f(\mathtt{obs})[i] = \mathtt{obs}[t]$ if $\mathtt{obs}[i] = *$ and $\mathtt{obs}[t] \in \{0, 1\}$; $f(\mathtt{obs})[i] = \mathtt{obs}[i]$ if $\mathtt{obs}[i] \in \{0, 1\}$ and $\mathtt{obs}[t] \in \{0, 1\}$. For every probability distribution $\mathcal{D}$ supporting $\mathcal{C}$ for x_t, with c being the target concept for x_t under $\mathcal{D}$, and every masking process $\mathtt{mask}$, define the probability distribution $\mathcal{D}'$ to be equal to the induced probability distribution $f(\mathtt{mask}(\mathcal{D}))$, and define the masking process $\mathtt{mask}'$ to be the identity mapping.

We proceed to prove some properties of monotone formulas with respect to the instance mapping f. For every observation $\mathtt{obs} \in \{0, 1, *\}^{|\mathcal{A}|}$, and every attribute $x'_i \in \mathcal{A}'$, the definition of f directly implies that:

- if $\mathtt{val}(x_i \mid \mathtt{obs}) \in \{0, 1\}$ and $\mathtt{obs}[t] \in \{0, 1\}$, then $\mathtt{val}(x'_i \mid f(\mathtt{obs})) = \mathtt{val}(x_i \mid \mathtt{obs})$;
- if $\mathtt{val}(x_i \mid \mathtt{obs}) = *$ and $\mathtt{obs}[t] \in \{0, 1\}$, then $\mathtt{val}(x'_i \mid f(\mathtt{obs})) = \mathtt{obs}[t]$.

Thus, for every observation $\mathtt{obs} \in \{0, 1, *\}^{|\mathcal{A}|}$, and every formula $\varphi' \in \{x'_t\} \cup \mathcal{C}' \cup \mathcal{H}'$, it holds that:

- if $\mathtt{val}(\varphi \mid \mathtt{obs}) \in \{0, 1\}$ and $\mathtt{obs}[t] \in \{0, 1\}$, then $\mathtt{val}(\varphi' \mid f(\mathtt{obs})) = \mathtt{val}(\varphi \mid \mathtt{obs})$;
- if $\mathtt{val}(\varphi \mid \mathtt{obs}) = *$ and $\mathtt{obs}[t] \in \{0, 1\}$, then $\mathtt{val}(\varphi' \mid f(\mathtt{obs})) = \mathtt{obs}[t]$.

For Condition (i) of Definition 4.2, consider a hypothesis $h' \in \mathcal{H}'$ and an observation $\mathtt{obs} \in \{0, 1, *\}^{|\mathcal{A}|}$ such that $g(h')$ has a consistency conflict with x_t w.r.t. $\mathtt{obs}$. Then, $\{\mathtt{val}(g(h') \mid \mathtt{obs}), \mathtt{obs}[t]\} = \{0, 1\}$. By the properties discussed above, it follows that $\mathtt{val}(h' \mid f(\mathtt{obs})) = \mathtt{val}(g(h') \mid \mathtt{obs})$ and $\mathtt{val}(x'_t \mid f(\mathtt{obs})) = \mathtt{obs}[t]$. Hence, $\{\mathtt{val}(h' \mid f(\mathtt{obs})), \mathtt{val}(x'_t \mid f(\mathtt{obs}))\} = \{0, 1\}$, and, therefore, h' has a consistency conflict with x'_t w.r.t. $f(\mathtt{obs})$, as needed.

Condition (ii) of Definition 4.2 follows trivially by definition of the probability distribution $\mathcal{D}'$, and the masking process $\mathtt{mask}'$. To establish that $\mathcal{D}'$ supports $\mathcal{C}'$ for x'_t, we show that the formula $c' = c \in \mathcal{C}'$ is the target concept for x'_t under $\mathcal{D}'$. Consider any fixed observation $\mathtt{obs}$ drawn from $\mathtt{mask}(\mathcal{D})$. Clearly, c does not have a consistency conflict with x_t w.r.t. $\mathtt{obs}$, and thus $\{\mathtt{val}(c' \mid \mathtt{obs}), \mathtt{obs}[t]\} = \{\mathtt{val}(c \mid \mathtt{obs}), \mathtt{obs}[t]\} \neq \{0, 1\}$. We proceed by case analysis on the remaining possibilities.

- In the first case assume that $\mathtt{obs}[t] = *$. By definition of the instance mapping f, it follows that $\mathtt{val}(x'_i \mid f(\mathtt{obs})) = 0$ for all $x'_i \in \mathcal{A}'$. Thus, any monotone formula φ other than the tautology is such that $\mathtt{val}(\varphi \mid f(\mathtt{obs})) = 0$. In particular, $\mathtt{val}(c' \mid f(\mathtt{obs})) = 0 = \mathtt{val}(x'_t \mid f(\mathtt{obs}))$.
- In the second case assume that $\mathtt{obs}[t] \in \{0, 1\}$ and $\mathtt{val}(c \mid \mathtt{obs}) = \mathtt{obs}[t]$. By the properties discussed earlier, it follows that $\mathtt{val}(c' \mid f(\mathtt{obs})) = \mathtt{val}(c \mid \mathtt{obs})$ and $\mathtt{val}(x'_t \mid f(\mathtt{obs})) = \mathtt{obs}[t]$. The assumption implies that $\mathtt{val}(c' \mid f(\mathtt{obs})) = \mathtt{val}(x'_t \mid f(\mathtt{obs}))$.
- In the third case assume that $\mathtt{obs}[t] \in \{0, 1\}$ and $\mathtt{val}(c \mid \mathtt{obs}) = *$. By the properties discussed earlier, it follows that $\mathtt{val}(c' \mid f(\mathtt{obs})) = \mathtt{obs}[t]$ and $\mathtt{val}(x'_t \mid f(\mathtt{obs})) = \mathtt{obs}[t]$. Therefore, $\mathtt{val}(c' \mid f(\mathtt{obs})) = \mathtt{val}(x'_t \mid f(\mathtt{obs}))$.

In each case we have established that $\mathtt{val}(c' \mid f(\mathtt{obs})) = \mathtt{val}(x'_t \mid f(\mathtt{obs}))$ for every observation $f(\mathtt{obs})$ drawn from $f(\mathtt{mask}(\mathcal{D}))$. Since observations drawn from $f(\mathtt{mask}(\mathcal{D}))$ are complete, and since $\mathcal{D}' = f(\mathtt{mask}(\mathcal{D}))$, we obtain that $\mathtt{val}(c' \mid \mathtt{exm}') = \mathtt{val}(x'_t \mid \mathtt{exm}')$ for every example $\mathtt{exm}'$ drawn from $\mathcal{D}'$. Definition 2.1 implies that x'_t is expressed by c' w.r.t. $\mathcal{D}'$; thus, c' is the target concept for x'_t under $\mathcal{D}'$.

With regards to Condition (iii) of Definition 4.2, the instance mapping f and the hypothesis mapping g are clearly computable in time linear in the size of their inputs, the number r of resulting learning tasks is 1, and *size*(c') is equal to *size*(c). At this point the reduction has been established.

The totality of the established reduction follows by Definition 4.3 and by definition of the instance mapping f. This concludes the proof. □

Theorem 4.3 establishes that the monotonicity of formulas in the concept and hypothesis classes is a *sufficient condition* under which the lack of complete information does not affect learnability. Interestingly enough, consistent learnability from partial observations reduces to consistent learnability of the *same* concept class from complete observations. Equally intriguing is the fact that a hypothesis learned (from complete observations) in the resulting learning task, applies *unmodified* for making predictions (on partial observations) in the original learning task. Since this same hypothesis is appropriate also for information recovery (cf. Theorem 2.2), it follows that a concrete strategy to accurately recover missing information is to simply assign appropriate default truth-values to masked attributes during the learning phase, consistently learn from the resulting complete observations, and employ the learned hypothesis *as is* to make predictions.

Two technical points are worth discussing here. The first one relates to the requirement for the tautology formula not to be part of the concept class. This restriction is without loss of generality. An agent attempting to learn the structure of its environment may always employ sampling to determine, with high probability, whether the target concept for a given target attribute could be the tautology, and employ Theorem 4.3 only when this is not the case. The second technical point relates to the encoding of the value of the target attribute in certain attributes of the resulting learning task. Although agnostic to this fact, an agent learning in the resulting learning task utilizes the value of the target attribute in a much more involved manner than its typical use as a means to test the predictions of hypotheses. What makes the established result non-trivial, is the fact that the returned hypothesis does not depend on the target attribute in the context of the original learning task, in which the hypothesis is eventually employed for making predictions.

A useful sufficient condition for consistent learnability follows immediately by Theorems 4.2 and 4.3.

Corollary 4.4 *(Sufficient condition for consistent learnability). Consider a learning task $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ over $\mathcal{A}$. The concept class $\mathcal{C}$ is consistently learnable on x_t by $\mathcal{H}$, if the concept class $\mathcal{C}$ is learnable by $\mathcal{H}$ under the Probably Approximately Correct semantics, and both $\mathcal{C}$ and $\mathcal{H}$ are classes of monotone formulas.*

Building on known learnability results under the PAC semantics [5,14,28], Corollary 4.4 implies learnability results for certain concept classes under the consistent learnability semantics. *Distribution-specific* PAC learnability results — where learning is expected to succeed only for a particular (often the uniform) probability distribution — do not fall under the auspices of Corollary 4.4, since the reduction that transforms a learning task in the context of consistent learnability to one under the PAC semantics, distorts the probability distribution over examples.

Corollary 4.5 *(Proper consistent learnability of certain concept classes). Each of the concept classes in {conjunctions, disjunctions, k-CNF, k-DNF, linear thresholds} of monotone formulas over $\mathcal{A} \setminus \{x_t\}$, is properly consistently learnable on the target attribute $x_t \in \mathcal{A}$.*

4.4. Shallowness preserves learnability

One of the tools employed by humans in modelling their environment is that of abstraction. The same tool can be employed while learning. Structure captured by a complex formula can be abstracted into a monotone disjunction, with each disjunct representing a complex situation. In some of these cases, the ability to learn the latter type of formulas (cf. Corollary 4.5) might imply the ability to learn the former one — this is so when abstraction is applied with certain moderation. We develop next the notions necessary to model the process of abstraction during learning, and to determine the degree of moderation that preserves learnability.

In terms of a given formula φ over a set of attributes $\mathcal{A}$, abstraction may be thought of as the process of substituting *new* attributes for sub-formulas of φ, in a manner prescribed by a set $\mathcal{M}$ of ***substitutions***. Recall that we think of formulas as syntactic objects; equivalently, we may think of each formula as corresponding to a particular circuit that computes the formula. Abstraction, then, is the process of replacing parts of the representation of a formula (i.e., certain sub-circuits with their associated inputs) with new attributes. Each substitution in $\mathcal{M}$ is of the form $x'_{i(\psi)}/\psi$, and indicates that attribute $x'_{i(\psi)} \in \mathcal{A}'$ is to be substituted for the sub-formula ψ. We require that $\mathcal{M}$ induces a bijection from sub-formulas ψ to attributes $x'_{i(\psi)} \in \mathcal{A}'$, so that the substitution process is invertible, a property that is critical for the abstraction to make sense. In the general case, substitutions may be applied non-deterministically on a formula φ, and more than one possible resulting formula may be produced. The unique maximal subset of all such resulting formulas that obeys the following constraints is known as the ***basis of*** φ ***given*** $\mathcal{M}$, and is denoted by $\mathtt{basis}(\varphi \mid \mathcal{M})$:

(i) for every formula $\varphi' \in \mathtt{basis}(\varphi \mid \mathcal{M})$, and every pair ψ_1, ψ_2 of sub-formulas of φ that belong in the set $\{\psi \mid x'_{i(\psi)}/\psi \in \mathcal{M}; x'_{i(\psi)} \text{ appears in } \varphi'\}$, there is no attribute $x_i \in \mathcal{A}$ that is shared by ψ_1 and ψ_2;
(ii) each formula $\varphi' \in \mathtt{basis}(\varphi \mid \mathcal{M})$ is over the new set of attributes $\mathcal{A}' = \{x'_{i(\psi)} \mid x'_{i(\psi)}/\psi \in \mathcal{M}\}$.

Roughly speaking, Condition (i) asks that the abstracted components are independent of each other, a restriction imposed to ensure that learnability is preserved, while Condition (ii) asks that all attributes in φ are replaced during the substitution process, a restriction imposed for notational convenience. Note that Condition (i) is trivially satisfied for a *read-once* formula φ, in which each attribute appears at most once. A number of valid and invalid sets of substitutions are illustrated in Table 2. An intuitive graphical illustration of the substitution process and the constraints it is defined to respect is depicted in Fig. 2.

Definition 4.4 *(Shallowness in classes of formulas).* A class $\mathcal{F}$ of formulas over a set of attributes $\mathcal{A}$ is ***shallow for*** a class $\mathcal{F}'$ of formulas over set of attributes $\mathcal{A}'$ ***w.r.t.*** a set $\mathcal{M}$ of substitutions, if $\mathcal{F}'$ is a subset of $\bigcup_{\varphi \in \mathcal{F}} \mathtt{basis}(\varphi \mid \mathcal{M})$ such that for each formula $\varphi \in \mathcal{F}$, there exists a formula $\varphi' \in \mathtt{basis}(\varphi \mid \mathcal{M}) \cap \mathcal{F}'$.

A class $\mathcal{F}$ formulas that is shallow for a class $\mathcal{F}'$ of formulas w.r.t. a set $\mathcal{M}$ of substitutions, contains formulas that exhibit structure not fundamentally different (as determined by $\mathcal{M}$) from the structure exhibited by formulas in $\mathcal{F}'$ (cf. Table 2). Thus, any given class of read-once formulas is shallow for the *same* class of monotone formulas w.r.t. the set of

Table 2
The bases of the formula $\overline{(x_{80} \rightarrow x_5)} \vee (\overline{x_7} \wedge x_2) \vee (x_9 \oplus \overline{x_{56}})$, given various sets of substitutions. Whenever a set of substitutions is valid, the formula's underlying disjunctive nature is preserved in its basis.

Set $\mathcal{M}$ of substitutions	Basis of formula given $\mathcal{M}$
$\{x'_1/\overline{(x_{80} \rightarrow x_5)},\ x'_2/(\overline{x_7} \wedge x_2),\ x'_3/(x_9 \oplus \overline{x_{56}})\}$	$\{x'_1 \vee x'_2 \vee x'_3\}$
$\{x'_1/\overline{(x_{80} \rightarrow x_5)},\ x'_2/(\overline{x_7} \wedge x_2),\ x'_2/(x_9 \oplus \overline{x_{56}})\}$	invalid $\mathcal{M}$: x'_2 not invertible
$\{x'_1/\overline{(x_{80} \rightarrow x_5)},\ x'_2/(\overline{x_7} \wedge x_2),\ x'_3/(x_9 \oplus \overline{x_{56}}),\ x'_4/(x_{80} \rightarrow x_5)\}$	$\{x'_1 \vee x'_2 \vee x'_3,\ \overline{x'_4} \vee x'_2 \vee x'_3\}$
$\{x'_1/\overline{(x_{80} \rightarrow x_5)} \vee (x_9 \oplus \overline{x_{56}}),\ x'_2/(\overline{x_7} \wedge x_2)\}$	$\{x'_1 \vee x'_2\}$
$\{x'_1/(x_{80} \rightarrow x_5),\ x'_2/x_7,\ x'_3/x_2,\ x'_4/x_9,\ x'_5/\overline{x_{56}}\}$	$\{\overline{x'_1} \vee (\overline{x'_2} \wedge x'_3) \vee (x'_4 \oplus x'_5)\}$
$\{x'_1/(x_{80} \rightarrow x_5),\ x'_2/(\overline{x_7} \wedge x_2),\ x'_3/x_9\}$	$\{\}$ (x_{56} is not replaceable)

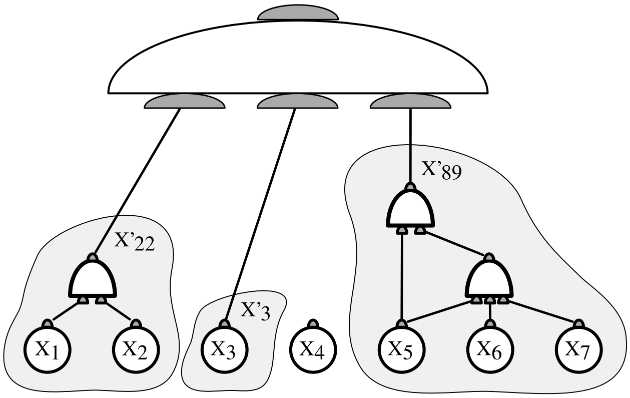

Fig. 2. Graphical illustration of the substitution process operating on a circuit that implements a formula φ over $\mathcal{A}$. Each substitution $x'_{i(\psi)}/\psi \in \mathcal{M}$ corresponds to a gadget that implements formula ψ, with $x'_{i(\psi)}$ standing for the name of that gadget. For notational convenience, we assume that for every attribute $x_i \in \mathcal{A}$, the substitution x'_i/x_i belongs in $\mathcal{M}$; hence, each circuit input x_i is implemented by gadget x'_i in $\mathcal{M}$. The substitution process amounts to employing gadgets from $\mathcal{M}$ to replace the shaded parts of the circuit. Once a circuit input has been replaced with a gadget, it becomes unavailable, so that other gadgets that refer to that input may no longer be employed. Nonetheless, gadgets may internally refer to the same input multiple times. When all the circuit inputs have been replaced with gadgets, we are left with a new circuit over new inputs that correspond to the names of the gadgets in $\mathcal{M}$. The resulting circuit may vary, depending on the choice of gadgets that were employed. Every new circuit is a truncated version, and hence an abstraction, of the original circuit, and the formula φ' that it implements is an element of the basis of φ given $\mathcal{M}$.

substitutions that replace each possible literal with a new attribute. Similarly, the class of read-once formulas in disjunctive normal form is shallow for the class of disjunctions w.r.t. the set of substitutions that replace each possible term (i.e., conjunction of literals) with a new attribute.

In the context of learning, $\mathcal{F}$ and $\mathcal{F}'$ may be viewed as representing the possible structures of two different environments; the structure of the first environment corresponds to some formula in $\mathcal{F}$, while the structure of the second environment corresponds to some formula in $\mathcal{F}'$. Establishing that $\mathcal{F}$ is shallow for $\mathcal{F}'$ w.r.t. $\mathcal{M}$, then, implies that the structure of the second environment is essentially an abstraction of the structure of the first one. As we have already pointed out, if the extent of this abstraction is moderate, then it might be possible to establish that learnability in the environment with the more abstract structure carries over to the environment with the more refined structure. The following definition describes conditions under which this is possible, in terms of $\mathcal{M}$, which is what ultimately determines the relation between $\mathcal{F}$ and $\mathcal{F}'$.

Definition 4.5 *(Moderately shallow learning tasks).* A learning task $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ over $\mathcal{A}$ is ***moderately shallow for*** a learning task $\langle x'_t, \mathcal{C}', \mathcal{H}'\rangle$ over $\mathcal{A}'$, if there exists a set $\mathcal{M}$ of substitutions such that:

(i) $\texttt{basis}(x_t \mid \mathcal{M}) = \{x'_t\}$, $\mathcal{C}$ is shallow for $\mathcal{C}'$ w.r.t. $\mathcal{M}$, and $\mathcal{H}$ is shallow for $\mathcal{H}'$ w.r.t. $\mathcal{M}$;
(ii) $\mathcal{M}$ is enumerable in time polynomial in $|\mathcal{A}|$;
(iii) for every $x'_{i(\psi)}/\psi \in \mathcal{M}$, and every observation $\texttt{obs} \in \{\texttt{0}, \texttt{1}, *\}^{|\mathcal{A}|}$, it holds that $\texttt{val}(\psi \mid \texttt{obs})$ is computable in time polynomial in $|\mathcal{A}|$.

Theorem 4.6 *(Reduction of moderately shallow learning tasks). A learning task $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ over $\mathcal{A}$ is reducible to a learning task $\langle x'_t, \mathcal{C}', \mathcal{H}'\rangle$ over $\mathcal{A}'$, if the former is moderately shallow for the latter.*

Proof. Assume that the learning task $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ over $\mathcal{A}$ is moderately shallow for the learning task $\langle x'_t, \mathcal{C}', \mathcal{H}'\rangle$ over $\mathcal{A}'$, and let $\mathcal{M}$ be the set of substitutions whose existence is guaranteed by Definition 4.5.

We first define the constructs whose existence is required by Definition 4.2. By Definition 4.4, for every hypothesis $h' \in \mathcal{H}'$, there exists a hypothesis $h \in \mathcal{H}$ such that $h' \in \texttt{basis}(h \mid \mathcal{M})$, and by definition of the set $\mathcal{M}$ of substitutions, h is unique; define the hypothesis mapping $g : \mathcal{H}' \to \mathcal{H}$ to map formula h' to this unique formula h. Define the instance mapping $f : \{0, 1, *\}^{|\mathcal{A}|} \to \{0, 1, *\}^{|\mathcal{A}'|}$ so that for every observation $\texttt{obs} \in \{0, 1, *\}^{|\mathcal{A}|}$, and every attribute $x'_{i(\psi)} \in \mathcal{A}'$, it holds that $f(\texttt{obs})[i(\psi)] = \texttt{val}(\psi \mid \texttt{obs})$; for every example $\texttt{exm}$ masked by $\texttt{obs}$, the definition of formula evaluation implies that $\texttt{val}(\psi \mid \texttt{obs}) \in \{\texttt{val}(\psi \mid \texttt{exm}), *\}$, which by definition of the instance mapping f implies that $f(\texttt{obs})[i(\psi)] \in \{f(\texttt{exm})[i(\psi)], *\}$, and thus that $f(\texttt{obs})$ masks $f(\texttt{exm})$. For every probability distribution $\mathcal{D}$ supporting $\mathcal{C}$ for x_t, with c being the target concept for x_t under $\mathcal{D}$, and every masking process $\texttt{mask}$, define the probability distribution $\mathcal{D}'$ to be equal to the induced probability distribution $f(\mathcal{D})$, and define the masking process $\texttt{mask}'$ so that for every example $\texttt{exm} \in \{0, 1\}^{|\mathcal{A}|}$, $\texttt{mask}'$ maps $f(\texttt{exm})$ to $f(\texttt{obs})$, where $\texttt{obs} \leftarrow \texttt{mask}(\texttt{exm})$; note that $f(\texttt{obs})$ masks $f(\texttt{exm})$, so $\texttt{mask}'$ is well-defined.

We proceed to prove some properties of formulas with respect to the instance mapping f. For every observation $\texttt{obs} \in \{0, 1, *\}^{|\mathcal{A}|}$, and every $x'_{i(\psi)}/\psi \in \mathcal{M}$, the definition of f directly implies that:

- $\texttt{val}(x'_{i(\psi)} \mid f(\texttt{obs})) = \texttt{val}(\psi \mid \texttt{obs})$.

Now, fix a formula $\varphi \in \{x_t\} \cup \mathcal{C} \cup \mathcal{H}$, a formula $\varphi' \in \texttt{basis}(\varphi \mid \mathcal{M})$, and an observation $\texttt{obs} \in \{0, 1, *\}^{|\mathcal{A}|}$. We continue to show how for each example $\texttt{exm}' \in \{0, 1\}^{|\mathcal{A}'|}$ that is masked by $f(\texttt{obs})$, one may construct an example $\texttt{exm} \in \{0, 1\}^{|\mathcal{A}|}$ that is masked by $\texttt{obs}$, and is such that $\texttt{val}(\varphi' \mid \texttt{exm}') = \texttt{val}(\varphi \mid \texttt{exm})$. The sought example $\texttt{exm}$ is obtained by starting from observation $\texttt{obs}$ and proceeding as follows:

> For every $x'_{i(\psi)}/\psi \in \mathcal{M}$ such that attribute $x'_{i(\psi)}$ appears in φ', and $\texttt{val}(\psi \mid \texttt{obs}) = *$, fix the masked attributes of $\texttt{obs}$ that appear in ψ so that ψ will evaluate to $\texttt{val}(x'_{i(\psi)} \mid \texttt{exm}')$. Fix any remaining masked attributes of $\texttt{obs}$ to any arbitrary $\{0, 1\}$ value to obtain $\texttt{exm}$.

By the requirement that the sub-formulas of φ that were replaced to obtain φ' do not share any attributes, it follows that the construction of example $\texttt{exm}$ is well-defined. It is also clear that $\texttt{exm}$ is masked by $\texttt{obs}$. Also, for each sub-formula ψ of φ that was replaced with an attribute $x'_{i(\psi)}$ it holds that $\texttt{val}(x'_{i(\psi)} \mid \texttt{exm}') = \texttt{val}(\psi \mid \texttt{exm})$. Indeed, either $\texttt{val}(\psi \mid \texttt{obs}) = *$, in which case the construction of $\texttt{exm}$ guarantees the claimed condition holds, or $\texttt{val}(\psi \mid \texttt{obs}) \in \{0, 1\}$, in which case the claimed condition follows, since: $\texttt{val}(\psi \mid \texttt{exm}) = \texttt{val}(\psi \mid \texttt{obs})$, since $\texttt{obs}$ masks $\texttt{exm}$; $\texttt{val}(x'_{i(\psi)} \mid f(\texttt{obs})) = \texttt{val}(\psi \mid \texttt{obs})$, by definition of the instance mapping f; $\texttt{val}(x'_{i(\psi)} \mid \texttt{exm}') = \texttt{val}(x'_{i(\psi)} \mid f(\texttt{obs}))$, since $f(\texttt{obs})$ masks $\texttt{exm}'$. Thus, for every example $\texttt{exm}' \in \{0, 1\}^{|\mathcal{A}'|}$ that is masked by $f(\texttt{obs})$, there exists an example $\texttt{exm} \in \{0, 1\}^{|\mathcal{A}|}$ that is masked by $\texttt{obs}$, such that $\texttt{val}(\varphi' \mid \texttt{exm}') = \texttt{val}(\varphi \mid \texttt{exm})$. This now implies that:

- for every formula $\varphi' \in \texttt{basis}(\varphi \mid \mathcal{M})$, $\texttt{val}(\varphi' \mid f(\texttt{obs})) = \texttt{val}(\varphi \mid \texttt{obs})$ if $\texttt{val}(\varphi \mid \texttt{obs}) \in \{0, 1\}$.

For Condition (i) of Definition 4.2, consider a hypothesis $h' \in \mathcal{H}'$ and an observation $\texttt{obs} \in \{0, 1, *\}^{|\mathcal{A}|}$ such that $g(h')$ has a consistency conflict with x_t w.r.t. $\texttt{obs}$. Then, $\{\texttt{val}(g(h') \mid \texttt{obs}), \texttt{obs}[t]\} = \{0, 1\}$. By definition of the hypothesis mapping, $g(h') \in \mathcal{H}$ is such that $h' \in \texttt{basis}(g(h') \mid \mathcal{M})$. By the properties discussed above, it follows that $\texttt{val}(h' \mid f(\texttt{obs})) = \texttt{val}(g(h') \mid \texttt{obs})$ and $\texttt{val}(x'_t \mid f(\texttt{obs})) = \texttt{obs}[t]$. Hence, $\{\texttt{val}(h' \mid f(\texttt{obs})), \texttt{val}(x'_t \mid f(\texttt{obs}))\} = \{0, 1\}$, and, therefore, h' has a consistency conflict with x'_t w.r.t. $f(\texttt{obs})$, as needed.

Condition (ii) of Definition 4.2 follows immediately by definition of the probability distribution $\mathcal{D}'$, and the masking process $\texttt{mask}'$, since $\mathcal{D}' = f(\mathcal{D})$, and $\texttt{mask}'(f(\mathcal{D})) = f(\texttt{mask}(\mathcal{D}))$. To establish that $\mathcal{D}'$ supports $\mathcal{C}'$ for x'_t, we show that any formula $c' \in \texttt{basis}(c \mid \mathcal{M}) \cap \mathcal{C}'$ is the target concept for x'_t under $\mathcal{D}'$; by Definition 4.4, such a formula exists. Since c is the target concept for x_t under $\mathcal{D}'$, Definition 2.1 implies that x_t is expressed by c w.r.t. $\mathcal{D}$, and thus

$$Pr\big[\texttt{val}(c \mid \texttt{exm}) = \texttt{exm}[t] \mid \texttt{exm} \leftarrow \mathcal{D}\big] = 1.$$

By the properties discussed above, it follows that $\texttt{val}(c' \mid f(\texttt{exm})) = \texttt{val}(c \mid \texttt{exm})$ and $\texttt{val}(x'_t \mid f(\texttt{exm})) = \texttt{exm}[t]$, for every example $\texttt{exm} \in \{0, 1\}^{|\mathcal{A}|}$. Hence,

$$Pr\big[\texttt{val}\big(c' \mid f(\texttt{exm})\big) = \texttt{val}\big(x'_t \mid f(\texttt{exm})\big) \mid \texttt{exm} \leftarrow \mathcal{D}\big] = 1.$$

Since it also holds that $\mathcal{D}' = f(\mathcal{D})$, we conclude that

$$Pr\big[\texttt{val}\big(c' \mid \texttt{exm}'\big) = \texttt{val}\big(x'_t \mid \texttt{exm}'\big) \mid \texttt{exm}' \leftarrow \mathcal{D}'\big] = 1.$$

Definition 2.1 implies that c' is the target concept for x'_t under $\mathcal{D}'$.

With regards to Condition (iii) of Definition 4.2, the instance mapping f is computable in the time required to traverse the set $\mathcal{M}$ of substitutions, and evaluate each of the associated sub-formulas on an observation; by Conditions (ii) and (iii) of Definition 4.5, both operations can be carried out in time polynomial in $|\mathcal{A}|$. The hypothesis mapping g is computable in the time required to read its input, and traverse the set $\mathcal{M}$ of substitutions to identify the sub-formula to be substituted for each attribute in the input formula; by Condition (ii) of Definition 4.5, each sub-formula can be identified in time polynomial in $|\mathcal{A}|$. The number r of resulting learning tasks is 1, and $size(c')$ is at most equal to $size(c)$. At this point the reduction has been established, and the proof is complete. □

A generalized version of the sufficient condition for consistent learnability that was established by Corollary 4.4 follows when Corollary 4.4 is taken in conjunction with Theorems 4.2 and 4.6.

Corollary 4.7 *(Generalized sufficient condition for consistent learnability). Consider a learning task $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ over $\mathcal{A}$. The concept class $\mathcal{C}$ is consistently learnable on x_t by $\mathcal{H}$, if the learning task $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ over $\mathcal{A}$ is moderately shallow for the learning task $\langle x'_t, \mathcal{C}', \mathcal{H}'\rangle$ over $\mathcal{A}'$, $\mathcal{C}'$ is learnable by $\mathcal{H}'$ under the Probably Approximately Correct semantics, and both $\mathcal{C}'$ and $\mathcal{H}'$ are classes of monotone formulas.*

This generalized sufficient condition implies the consistent learnability of additional concept classes. For the classes of conjunctions, disjunctions, and linear thresholds, the following result generalizes Corollary 4.5, by retracting the monotonicity assumption. For the classes of k-CNF and k-DNF formulas, the following result provides new consistently learnable subclasses that are incomparable to the subclasses whose consistent learnability was established by Corollary 4.5, by substituting the read-once property for the monotonicity property; k-CNF and k-DNF formulas may have none, either, or both of these two properties. As for Corollary 4.4, distribution-specific PAC learnability results do not fall under the auspices of Corollary 4.7.

Corollary 4.8 *(Proper consistent learnability of additional concept classes). Each of the concept classes in {conjunctions, disjunctions, read-once k-CNF, read-once k-DNF, linear thresholds} of formulas over literals in $\mathcal{A} \setminus \{x_t\}$, is properly consistently learnable on the target attribute $x_t \in \mathcal{A}$.*

In a preliminary version of this work it was incorrectly reported that the general classes of k-CNF and k-DNF formulas are properly consistently learnable. We find it informative to discuss the subtle, but critical, point that prevents our results from generalizing to these classes. Consider the formula $\varphi_1 \vee \varphi_2$. When this formula is evaluated on an example $\mathtt{exm}$, by definition it holds that $\mathtt{val}(\varphi_1 \vee \varphi_2 \mid \mathtt{exm}) = \mathtt{val}(\varphi_1 \mid \mathtt{exm}) \vee \mathtt{val}(\varphi_2 \mid \mathtt{exm})$; similar properties hold for other logical connectives. Observe, however, that such *local evaluation* of the formula cannot be carried out on partial observations. Indeed, if $\mathtt{val}(\varphi_1 \mid \mathtt{obs}) = \mathtt{val}(\varphi_2 \mid \mathtt{obs}) = *$, then $\mathtt{val}(\varphi_1 \vee \varphi_2 \mid \mathtt{obs})$ cannot, in general, be uniquely determined. If, for instance, φ_1 is semantically the negation of φ_2, then $\mathtt{val}(\varphi_1 \vee \varphi_2 \mid \mathtt{obs}) = \mathtt{1}$, whereas if φ_1 and φ_2 are semantically equivalent, then $\mathtt{val}(\varphi_1 \vee \varphi_2 \mid \mathtt{obs}) = *$. Note that the locality of evaluation is restored if the formulas φ_1 and φ_2 are assumed not to share any attributes. Although we are unaware of any relevant formal result in the learning literature, it seems natural to conjecture that locality of formula evaluation is essential for reductions to go through in learning settings. This, in turn, explains why while reductions can establish the learnability of general k-CNF and k-DNF formulas under the PAC semantics, they seem to be able to establish the learnability of only their read-once counterparts under the autodidactic learning semantics.

5. Negative learnability results

We have already pointed out that learnability under the PAC semantics is a special case of autodidactic learnability. So far, we have not excluded the possibility that the two learning models are equivalent in terms of the concept classes that are learnable. On the contrary, our general positive learnability results indicate that the two models are equivalent on a broad set of concept classes, tempting one to conjecture that lack of information during learning does not render learnability any harder — this we have shown to be true, for instance, for concept classes of monotone formulas. In this section we make some progress towards disproving such a conjecture. We show that two particular concept classes that are properly learnable under the PAC semantics are not *properly* learnable under the autodidactic learning semantics (i.e., if one insists that the concept and hypothesis classes coincide). Such representation-specific non-learnability results have been studied before in the context of PAC learnability [22], and do not preclude the possibility that such concept classes are non-properly learnable. The non-proper learnability under the autodidactic learning semantics of the two concept classes discussed in this section remains open.

The negative results that we prove are with respect to consistent learnability. Note that accuracy implies consistency, irrespectively of the concealment degree of the masking process. Thus, our results also imply that learning accurately is not possible in certain cases. It is worth emphasizing that the established negative results do not require the use of masking processes with a high degree of concealment. Indeed, observations with a masked target attribute do not constrain the learner in any way (since any hypothesis makes consistent predictions on such observations), and thus using a masking process that masks the target attribute in any observation does not offer any advantage. All our results employ only 0-concealing masking processes. We also note that our results *do not* rely on using formulas that cannot be efficiently evaluated on observations.

We start with a general result that we later use to obtain specific negative autodidactic learnability results.

Theorem 5.1 *(Sufficient condition for hard learning tasks). Fix an arbitrary positive integer $n \in \mathbb{N}$. Consider a set of attributes $\mathcal{A}$ of size polynomial in n, and a learning task $\langle x_t, \mathcal{C}, \mathcal{H} \rangle$ over $\mathcal{A}$ such that $\log|\mathcal{C}|$ is of size polynomial in n. Assume that there exists an algorithm that on input a 3-CNF formula χ of size n runs in time polynomial in n and outputs a set of observations $\mathcal{O}(\chi) \subseteq \{0, 1, *\}^{|\mathcal{A}|}$ such that*:

(i) *every formula in $\mathcal{H}$ is evaluatable in time polynomial in n on every observation in $\mathcal{O}(\chi)$;*
(ii) *χ is satisfiable if there exists a formula in $\mathcal{H}$ that is 1-consistent with x_t given $\mathcal{O}(\chi)$;*
(iii) *χ is satisfiable only if there exists a probability distribution $\mathcal{D}$ over $\{0, 1\}^{|\mathcal{A}|}$, and a masking process* `mask` *such that $\mathcal{D}$ supports $\mathcal{C}$ for x_t, and* `mask`$(\mathcal{D})$ *is the uniform distribution over $\mathcal{O}(\chi)$.*

Then, the concept class $\mathcal{C}$ is not consistently learnable on the target attribute x_t by the hypothesis class $\mathcal{H}$, unless RP = NP.

Proof. Assume that the concept class $\mathcal{C}$ is consistently learnable on the target attribute x_t by the hypothesis class $\mathcal{H}$. Let $\mathcal{L}$ be a consistent learner for the learning task $\langle x_t, \mathcal{C}, \mathcal{H} \rangle$ over $\mathcal{A}$, and let $q(\cdot, \cdot, \cdot, \cdot)$ be the associated polynomial that determines the running time of algorithm $\mathcal{L}$ given its input parameters. Consider the algorithm $\mathcal{L}_{sat}$ defined as follows:

> On input a 3-CNF formula χ of size n, algorithm $\mathcal{L}_{sat}$ constructs the set of observations $\mathcal{O}(\chi)$. It then proceeds to simulate algorithm $\mathcal{L}$ with input $\mathcal{A}$, $\langle x_t, \mathcal{C}, \mathcal{H} \rangle$, $\delta = 1/3$, $\varepsilon = 1/2|\mathcal{O}(\chi)|$, and an oracle returning observations. The simulation is interrupted after $q(1/\delta, 1/\varepsilon, |\mathcal{A}|, \log|\mathcal{C}|)$ time-steps. During the simulation, whenever algorithm $\mathcal{L}$ accesses the oracle and requests an observation, algorithm $\mathcal{L}_{sat}$ draws an observation `obs` uniformly at random from $\mathcal{O}(\chi)$, and passes `obs` to algorithm $\mathcal{L}$. If the simulated algorithm $\mathcal{L}$ returns a hypothesis $h \in \mathcal{H}$, algorithm $\mathcal{L}_{sat}$ checks and returns whether h does not have a consistency conflict with x_t w.r.t. any observation in $\mathcal{O}(\chi)$. If the simulation of algorithm $\mathcal{L}$ is interrupted, algorithm $\mathcal{L}_{sat}$ returns `false`. In either case, algorithm $\mathcal{L}_{sat}$ terminates after returning a truth-value.

We now prove that algorithm $\mathcal{L}_{sat}$ runs in time polynomial in n and determines whether a given arbitrary 3-CNF formula χ of size n is satisfiable so that: if χ is unsatisfiable, then algorithm $\mathcal{L}_{sat}$ will return `false` with probability 1; if χ is satisfiable, then algorithm $\mathcal{L}_{sat}$ will return `true` with probability at least 2/3.

Assume first that χ is unsatisfiable. By Condition (ii), every formula in $\mathcal{H}$ has a consistency conflict with x_t w.r.t. some observation in $\mathcal{O}(\chi)$. Therefore, algorithm $\mathcal{L}_{sat}$ will return `false`, as expected, irrespectively of whether algorithm $\mathcal{L}$ returns some hypothesis $h \in \mathcal{H}$, or its simulation is interrupted.

Assume now that χ is satisfiable. By Condition (iii), there exists a probability distribution $\mathcal{D}$ and a masking process `mask` such that $\mathcal{D}$ supports $\mathcal{C}$ for x_t, and the oracle of algorithm $\mathcal{L}$ draws observations from `mask`$(\mathcal{D})$; let c be the target concept for x_t under $\mathcal{D}$. By Definition 2.3, algorithm $\mathcal{L}$ will run in time $q(1/\delta, 1/\varepsilon, |\mathcal{A}|, size(c))$, and return, with probability $1 - \delta$, a hypothesis $h \in \mathcal{H}$ that is $(1 - \varepsilon)$-consistent with x_t under $\mathcal{D}$ and `mask`. Since $size(c) \leqslant \log|\mathcal{C}|$, then the simulation of algorithm $\mathcal{L}$ will not be interrupted, and algorithm $\mathcal{L}_{sat}$ will obtain h. Since ε is strictly less than the probability with which any particular observation from $\mathcal{O}(\chi)$ is drawn, it follows that, with probability $1 - \delta$, the returned hypothesis $h \in \mathcal{H}$ will have no consistency conflict with x_t w.r.t. any observation in $\mathcal{O}(\chi)$, and algorithm $\mathcal{L}_{sat}$ will verify this and return `true`, as expected. Since $\delta = 1/3$, the probability with which algorithm $\mathcal{L}_{sat}$ will correctly report that χ is satisfiable is at least 2/3.

Since $1/\delta = 3$, $1/\varepsilon = 2|\mathcal{O}(\chi)| = \text{poly}(n)$, $|\mathcal{A}| = \text{poly}(n)$, and $\log|\mathcal{C}| = \text{poly}(n)$, it follows that $q(1/\delta, 1/\varepsilon, |\mathcal{A}|, \log|\mathcal{C}|)$ is polynomial in n. The set of observations $\mathcal{O}(\chi)$ is constructible in time polynomial in n, observations are uniformly samplable from $\mathcal{O}(\chi)$ in time polynomial in n, and, by Condition (i), returned hypotheses are testable for consistency conflicts with x_t w.r.t. observations in $\mathcal{O}(\chi)$ in time polynomial in n. Hence, algorithm $\mathcal{L}_{sat}$ runs in time polynomial in n. In conclusion, we have established the existence of an algorithm, namely algorithm $\mathcal{L}_{sat}$, that solves an NP-complete problem within the resource constraints allowed for problems in RP. This implies that RP = NP, and concludes the proof. □

Under the standard computational complexity assumption that RP ≠ NP, we present next intractability results on the proper consistent learnability of certain explicit concept classes that are known to be properly PAC learnable. Our results hold even if *at most three* attributes are masked in observations, none of which is the target attribute, suggesting that the property of an agent's sensing process that compromises consistent learnability is not the frequency with which information is missing in the obtained appearances, but rather the context in which this happens.[3] This realization is further corroborated when viewed in conjunction with, and in contrast to, certain results from the literature that establish that learnability is not severely impaired when information in observations is missing independently at random on each attribute, despite this giving rise to observations with possibly *many* simultaneously masked attributes [6].

[3] Recall, by Theorem 2.1, that a similar phenomenon occurs also when learned hypotheses are eventually employed by an agent for making predictions. The predictive accuracy of hypotheses, even highly consistent ones, is compromised in a manner that depends not only on the frequency with which sensors hide information, but mainly on the context in which this happens. Intriguing is also the fact that in obtaining the impossibility of learning highly accurate hypotheses it suffices for the target attribute to be masked, whereas in obtaining the intractability of learning highly consistent hypotheses it suffices for non-target attributes to be masked. Hence, masked "features" are, in some sense, associated with a more fundamental reason for unlearnability as compared to masked "labels" in learning instances.

5.1. Non-learnability of parities

A *parity formula* over a set of attributes $\mathcal{A}$ is a formula of the form $x_{i_1} \oplus \cdots \oplus x_{i_r}$, where $\oplus$ denotes the "exclusive or" binary operator. A parity formula evaluates to `1` on an example `exm` exactly when an odd number of the formula's attributes are assigned the value `1` in `exm`.

The concept class of parity formulas is one associated with numerous open problems in the learning literature. Nonetheless, the concept class is known to be properly learnable under the PAC semantics [7,11], albeit using techniques that rely critically on the availability of an explicit set of complete observations. In particular, the concept class of parity formulas is known to be *unconditionally* non-learnable in the representation-independent sense in the Statistical Query model [15], and non-evolvable [31] indicating the singularity of this concept class. A justification of its singularity may appeal to the extreme sensitivity of parity formulas on their attributes; independently of the values of the remaining attributes, the change of an attribute's value affects the value of a parity formula. It is this property that the following result exploits. Interestingly, parity formulas are highly non-monotone, which is in accordance with the consistent learnability of concept classes of monotone formulas.

Theorem 5.2 *(Intractability of proper consistent learnability of parities). The concept class $\mathcal{C}$ of parities over $\mathcal{A} \setminus \{x_t\}$ is not properly consistently learnable on the target attribute $x_t \in \mathcal{A}$, unless* `RP` = `NP`.

Proof. Fix an arbitrary positive integer $n \in \mathbb{N}$. Let $V = \{v_1, v_2, \ldots, v_n\}$ be the set of variables over which instances of `3-SAT` of size n are defined. Construct the learning task $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ over $\mathcal{A}$ as follows: Define $\mathcal{A} \triangleq \{x_i^+, x_i^- \mid v_i \in V\} \cup \{x_t\}$, $\mathcal{C}$ to be the set of all parities over $\mathcal{A} \setminus \{x_t\}$, and $\mathcal{H} \triangleq \mathcal{C}$.

For each variable $v_i \in V$, we define $a(v_i) \triangleq x_i^+$ and $a(\overline{v_i}) \triangleq x_i^-$; by construction, the mapping from the set of literals over V to the set of attributes $\mathcal{A} \setminus \{x_t\}$ is bijective. For every 3-CNF formula

$$\chi = \bigwedge_{j=1}^{m} (l_{j,1} \vee l_{j,2} \vee l_{j,3}),$$

where each $l_{j,k}$ is a literal over V, denote by $\mathcal{O}(\chi)$ the set of observations that contains exactly the following:

(i) for every i: $1 \leqslant i \leqslant n$, the observation $\mathtt{obs}_{var(i)}$ that assigns the value `1` to attributes x_i^+, x_i^-, the value `1` to attribute x_t, and the value `0` to attributes in $\mathcal{A} \setminus \{x_i^+, x_i^-, x_t\}$;
(ii) for every j: $1 \leqslant j \leqslant m$, the observation $\mathtt{obs}_{cls(j)}$ that assigns the value $*$ to attributes $a(l_{j,1}), a(l_{j,2}), a(l_{j,3})$, the value `1` to attribute x_t, and the value `0` to attributes in $\mathcal{A} \setminus \{a(l_{j,1}), a(l_{j,2}), a(l_{j,3}), x_t\}$.

We continue to establish that formula χ is satisfiable if (and only if) there exists a parity formula $\varphi \in \mathcal{H}$ that does not have a consistency conflict with the target attribute x_t w.r.t. any observation in $\mathcal{O}(\chi)$. Consider any set of literals τ over V, and let $\varphi_\tau \triangleq \bigoplus\{a(l) \mid l \in \tau\} \in \mathcal{H}$ be the corresponding parity formula; by the bijective property of a, the mapping from the set of literal-sets over V to the set $\mathcal{H}$ of hypotheses is bijective. We next prove certain properties of this mapping. The following derivation establishes a correspondence between truth-assignments induced by τ for χ, and the lack of consistency conflicts of φ_τ with x_t w.r.t. the observations in $\mathcal{O}(\chi)$ that are of type (i):

$$\begin{aligned}
&\tau \text{ induces a truth-assignment for } \chi \\
&\quad\Leftrightarrow \\
&\text{for every } i\text{: } 1 \leqslant i \leqslant n, \text{ exactly one of } v_i, \overline{v_i} \text{ belongs in } \tau \\
&\quad\Leftrightarrow \\
&\text{for every } i\text{: } 1 \leqslant i \leqslant n, \text{ exactly one of } x_i^+, x_i^- \text{ belongs in } \varphi_\tau \\
&\quad\Leftrightarrow \\
&\text{for every } i\text{: } 1 \leqslant i \leqslant n, \ \varphi_\tau \text{ does not have a consistency conflict with } x_t \text{ w.r.t. } \mathtt{obs}_{var(i)}.
\end{aligned}$$

A second derivation establishes a correspondence between sets of literals determined by τ that would satisfy χ, and the lack of consistency conflicts of φ_τ with x_t w.r.t. the observations in $\mathcal{O}(\chi)$ that are of type (ii):

$$\begin{aligned}
&\tau \text{ contains at least one literal from each clause of } \chi \\
&\quad\Leftrightarrow \\
&\text{for every } j\text{: } 1 \leqslant j \leqslant m, \text{ at least one of } l_{j,1}, l_{j,2}, l_{j,3} \text{ belongs in } \tau \\
&\quad\Leftrightarrow
\end{aligned}$$

for every j: $1 \leqslant j \leqslant m$, at least one of $a(l_{j,1}), a(l_{j,2}), a(l_{j,3})$ belongs in φ_τ

$\Leftrightarrow$

for every j: $1 \leqslant j \leqslant m$, $\mathtt{val}(\varphi_\tau \mid \mathtt{obs}_{cls(j)}) = *$

$\Leftrightarrow$

for every j: $1 \leqslant j \leqslant m$, φ_τ does not have a consistency conflict with x_t w.r.t. $\mathtt{obs}_{cls(j)}$.

Together, the two derivations imply that τ induces a satisfying truth-assignment for χ if and only if φ_τ does not have a consistency conflict with x_t w.r.t. any observation in $\mathcal{O}(\chi)$. This conclusion then, along with the bijection property of the mapping, leads to the following derivation, which establishes the claim:

χ is satisfiable

$\Leftrightarrow$

there exists a set of literals τ over V that induces a satisfying truth-assignment for χ

$\Leftrightarrow$

there exists τ such that φ_τ does not have a consistency conflict with x_t w.r.t. any observation in $\mathcal{O}(\chi)$

$\Leftrightarrow$

there exists $\varphi \in \mathcal{H}$ that does not have a consistency conflict with x_t w.r.t. any observation in $\mathcal{O}(\chi)$.

To conclude the proof it suffices to show that the conditions of Theorem 5.1 are satisfied. Clearly, $|\mathcal{A}|$ is polynomial in n, and so is $\log|\mathcal{C}|$, since there are at most $2^{|\mathcal{A}|}$ formulas in $\mathcal{C}$. Also, the set of observations $\mathcal{O}(\chi)$ is constructible in time polynomial in n. For Condition (i) of Theorem 5.1, note that each parity $\varphi \in \mathcal{H}$ can be evaluated on each observation $\mathtt{obs} \in \mathcal{O}(\chi)$ in time polynomial in n. Indeed, either an attribute in φ is masked in $\mathtt{obs}$, in which case $\mathtt{val}(\varphi \mid \mathtt{obs}) = *$, or none of the attributes in φ is masked in $\mathtt{obs}$, in which case the number of attributes that are assigned the value $\mathtt{1}$ in $\mathtt{obs}$ determines $\mathtt{val}(\varphi \mid \mathtt{obs})$.

Condition (ii) of Theorem 5.1 follows directly from the last derivation above. For Condition (iii) of Theorem 5.1, assume that χ is satisfiable, and let τ be a set of literals over V that induces a satisfying truth-assignment for χ. Consider the set of examples that contains exactly the following:

(i) for every i: $1 \leqslant i \leqslant n$, the example $\mathtt{exm}_{var(i)}$ that assigns the value $\mathtt{1}$ to attributes x_i^+, x_i^-, the value $\mathtt{1}$ to attribute x_t, and the value $\mathtt{0}$ to attributes in $\mathcal{A} \setminus \{x_i^+, x_i^-, x_t\}$;
(ii) for every j: $1 \leqslant j \leqslant m$, the example $\mathtt{exm}_{cls(j)}$ that assigns the value $\mathtt{1}$ to the least (under some fixed ordering of the attributes in $\mathcal{A}$) attribute $a(l_{j,k})$ in the set $\{a(l_{j,1}), a(l_{j,2}), a(l_{j,3})\} \cap \{a(l) \mid l \in \tau\}$, the value $\mathtt{1}$ to attribute x_t, and the value $\mathtt{0}$ to attributes in $\mathcal{A} \setminus \{a(l_{j,k}), x_t\}$.

Define the probability distribution $\mathcal{D}$ that returns each of these examples with equal probability. Define the masking process $\mathtt{mask}$ that maps each example to the observation in $\mathcal{O}(\chi)$ with the same subscript with probability 1. Clearly, $\mathtt{mask}(\mathcal{D})$ is uniform over $\mathcal{O}(\chi)$, and the target attribute x_t is expressed by the parity formula $\varphi_\tau \in \mathcal{C}$ w.r.t. the probability distribution $\mathcal{D}$; thus, $\mathcal{D}$ supports $\mathcal{C}$ for x_t, and φ_τ is the target concept for x_t under $\mathcal{D}$. Condition (iii) of Theorem 5.1 follows, and the proof is complete. □

When compared with related results from the literature (see, e.g., [22]), the existence of partial observations complicates the required reduction from an NP-complete problem that lies in the heart of the proof of Theorem 5.2. Yet, partial observations also allow for a more flexible manipulation of the learning algorithm in the proof of Theorem 5.1 that is invoked in the reduction, which then explains the ability to establish the particular intractability result. More precisely, the reduction relies on constructing observations that in order to be explained consistently, require the learned hypothesis to depend on *any* non-empty subset of the *masked* attributes, without, however, the observations specifying which such subset is to be chosen. Such constraints allude to a combinatorial problem, which is precisely what the learning algorithm is expected to solve. It is the case that with complete observations one may still force the learned hypothesis to depend on certain subsets of attributes, but the possible dependencies on these attributes are necessarily restricted by the observations themselves in what seems to be a subtle, yet critical, manner.

5.2. Non-learnability of decision lists

A *k-decision list* over a set of attributes $\mathcal{A}$ is an *ordered* sequence $\langle c_1, v_1\rangle \ldots \langle c_r, v_r\rangle\langle \top, v_{r+1}\rangle$ of pairs comprising a condition c_i that is a term of at most $k \in \mathbb{N}$ literals over $\mathcal{A}$, and an associated decision v_i that is a $\{\mathtt{0}, \mathtt{1}\}$ value. A decision list evaluates on an example $\mathtt{exm}$ to the value v_i associated with the least-indexed condition c_i that evaluates to $\mathtt{1}$ on $\mathtt{exm}$; the tautology formula that appears as the last condition ensures that the evaluation process is well-defined.

The concept class of k-decision lists, for any *constant* $k \in \mathbb{N}$, is known to be properly learnable under the PAC semantics [25]. The proper learnability is retained even if only monotone-term k-decision lists are considered. Unlike parity formulas, (monotone-term) k-decision lists are learnable under the Statistical Query semantics [15], and in the presence of random classification noise [1].

Rivest [25], who introduced this concept class and established its PAC learnability, asked whether learnability is preserved when instead of complete observations one considers *partial* observations. In our notation, he defined *agreement*[4] of a formula φ with an observation $\texttt{obs}$ to mean $\texttt{val}(\varphi \mid \texttt{obs}) = \texttt{obs}[t]$, assuming that the target attribute x_t is never masked in drawn observations. As posed, the question almost always admits a trivial negative answer: an observation $\texttt{obs}$ generally masks a set of examples across which the value of φ varies, implying that $\texttt{val}(\varphi \mid \texttt{obs}) = *$, and making φ “disagree” with $\texttt{obs}$. We recast the notion of “agreement” to what, we believe, is a more appropriate (and possibly the intended) form: a formula φ *agrees* with an observation $\texttt{obs}$ if φ does not have a consistency conflict with the target attribute x_t under $\texttt{obs}$. This notion of “agreement” is weaker, as it requires that $\texttt{val}(\varphi \mid \texttt{obs}) = \texttt{obs}[t]$ only when $\texttt{val}(\varphi \mid \texttt{obs}) \in \{\texttt{0}, \texttt{1}\}$ and $\texttt{obs}[t] \in \{\texttt{0}, \texttt{1}\}$. We partially answer this new question in the negative, by showing the concept class of monotone-term 1-decision lists not to be *properly* consistently learnable from partial observations, unless $\texttt{RP} = \texttt{NP}$. The negative answer carries to Rivest’s original question, due to the stronger notion of “agreement” that he used.

Theorem 5.3 *(Intractability of proper consistent learnability of monotone-term* 1*-decision lists). The concept class $\mathcal{C}$ of monotone-term* 1*-decision lists over $\mathcal{A} \setminus \{x_t\}$ is not properly consistently learnable on the target attribute $x_t \in \mathcal{A}$, unless* $\texttt{RP} = \texttt{NP}$.

Proof. Fix an arbitrary positive integer $n \in \mathbb{N}$. Let $V = \{v_1, v_2, \ldots, v_n\}$ be the set of variables over which instances of $\texttt{3-SAT}$ of size n are defined. Construct the learning task $\langle x_t, \mathcal{C}, \mathcal{H} \rangle$ over $\mathcal{A}$ as follows: Define $\mathcal{A} \triangleq \{x_i^+, x_i^- \mid v_i \in V\} \cup \{x_t\}$, $\mathcal{C}$ to be the set of all monotone-term 1-decision lists over $\mathcal{A} \setminus \{x_t\}$, and $\mathcal{H} \triangleq \mathcal{C}$.

For each variable $v_i \in V$, we define $a(v_i) \triangleq x_i^+$ and $a(\overline{v_i}) \triangleq x_i^-$; by construction, the mapping from the set of literals over V to the set of attributes $\mathcal{A} \setminus \{x_t\}$ is bijective. For every 3-CNF formula

$$\chi = \bigwedge_{j=1}^{m} (l_{j,1} \vee l_{j,2} \vee l_{j,3}),$$

where each $l_{j,k}$ is a literal over V, denote by $\mathcal{O}(\chi)$ the set of observations that contains exactly the following:

(i) the observation $\texttt{obs}_{zero}$ that assigns the value $\texttt{0}$ to attributes in $\mathcal{A}$;
(ii) for every i: $1 \leqslant i \leqslant n$, the observation $\texttt{obs}_{var(i,\texttt{0})}$ that assigns the value $\texttt{1}$ to attributes x_i^+, x_i^-, the value $\texttt{0}$ to attribute x_t, and the value $\texttt{0}$ to attributes in $\mathcal{A} \setminus \{x_i^+, x_i^-, x_t\}$;
(iii) for every i: $1 \leqslant i \leqslant n$, the observation $\texttt{obs}_{var(i,\texttt{1})}$ that assigns the value $*$ to attributes x_i^+, x_i^-, the value $\texttt{1}$ to attribute x_t, and the value $\texttt{0}$ to attributes in $\mathcal{A} \setminus \{x_i^+, x_i^-, x_t\}$;
(iv) for every j: $1 \leqslant j \leqslant m$, the observation $\texttt{obs}_{cls(j)}$ that assigns the value $*$ to attributes $a(l_{j,1}), a(l_{j,2}), a(l_{j,3})$, the value $\texttt{1}$ to attribute x_t, and the value $\texttt{0}$ to attributes in $\mathcal{A} \setminus \{a(l_{j,1}), a(l_{j,2}), a(l_{j,3}), x_t\}$.

Without loss of generality, for the remainder of this proof we consider only *read-once* monotone-term 1-decision lists, where no attribute appears in conditions more than once; for any monotone-term 1-decision list that violates this assumption one may simply drop all but the first condition out of those that contain any particular attribute, and obtain a new monotone-term 1-decision list that respects the assumption and is equivalent to the first one.

We continue to establish that formula χ is satisfiable if there exists a monotone-term 1-decision list $\varphi \in \mathcal{H}$ that does not have a consistency conflict with the target attribute x_t w.r.t. any observation in $\mathcal{O}(\chi)$. Consider any monotone-term 1-decision list $\varphi \in \mathcal{H}$ that does not have a consistency conflict with x_t w.r.t. any observation in $\mathcal{O}(\chi)$, and let $\tau_\varphi \triangleq \{l \mid \langle a(l), \texttt{1} \rangle \in \varphi\}$ be a set of literals over V. We continue to verify that τ_φ induces a satisfying assignment for χ. For each observation $\texttt{obs} \in \mathcal{O}(\chi)$, we identify the conclusions that follow given that $\varphi \in \mathcal{H}$ does not have a consistency conflict with x_t w.r.t. $\texttt{obs}$. We proceed by case analysis on the four types of observations in $\mathcal{O}(\chi)$:

(i) from observation $\texttt{obs}_{zero}$, it follows that the value corresponding to the tautology condition of φ is $\texttt{0}$;
(ii) for every i: $1 \leqslant i \leqslant n$, from observation $\texttt{obs}_{var(i,\texttt{0})}$, it follows that φ does not contain both of the pairs $\langle x_i^+, \texttt{1} \rangle, \langle x_i^-, \texttt{1} \rangle$, for otherwise φ would evaluate to $\texttt{1}$ on $\texttt{obs}_{var(i,\texttt{0})}$;
(iii) for every i: $1 \leqslant i \leqslant n$, from observation $\texttt{obs}_{var(i,\texttt{1})}$, it follows that φ contains at least one of the pairs $\langle x_i^+, \texttt{1} \rangle, \langle x_i^-, \texttt{1} \rangle$, for otherwise only the tautology condition of φ would be satisfied, and by Conclusion (i), φ would evaluate to $\texttt{0}$ on $\texttt{obs}_{var(i,\texttt{1})}$;

[4] Rivest [25] actually used the term “consistency” in his work, rather than the term “agreement” that is employed here. We avoid, however, the use of the term “consistency” in this context, as this term has a different meaning in our framework.

(iv) for every j: $1 \leqslant j \leqslant m$, from observation $\mathtt{obs}_{cls(j)}$, it follows that φ contains at least one of the pairs $\langle a(l_{j,1}), \mathtt{1}\rangle$, $\langle a(l_{j,2}), \mathtt{1}\rangle$, $\langle a(l_{j,3}), \mathtt{1}\rangle$, for otherwise only the tautology condition of φ would be satisfied, and by Conclusion (i), φ would evaluate to $\mathtt{0}$ on $\mathtt{obs}_{cls(j)}$.

Conclusions (ii) and (iii) imply that τ_φ induces a truth-assignment for χ, and Conclusion (iv) implies that the induced truth-assignment is a satisfying one for χ; thus, χ is satisfiable, as needed.

To conclude the proof it suffices to show that the conditions of Theorem 5.1 are satisfied. Clearly, $|\mathcal{A}|$ is polynomial in n, and so is $\log|\mathcal{C}|$, since there are at most $2 \cdot 3^{|\mathcal{A}|}|\mathcal{A}|!$ formulas in $\mathcal{C}$. Also, the set of observations $\mathcal{O}(\chi)$ is constructible in time polynomial in n. For Condition (i) of Theorem 5.1, note that each monotone-term 1-decision list $\varphi \in \mathcal{H}$ can be evaluated on each observation $\mathtt{obs} \in \mathcal{O}(\chi)$ in time polynomial in n. Indeed, for each attribute in φ that is assigned the value $\mathtt{1}$ in $\mathtt{obs}$ one may prune the suffix of φ that follows the tuple $\langle c_i, v_i\rangle$ of φ whose condition c_i is the attribute, and replace the condition of the said tuple with the tautology formula $\top$, without affecting the value of φ on $\mathtt{obs}$. Similarly, for each attribute in φ that is assigned the value $\mathtt{0}$ in $\mathtt{obs}$ one may drop the tuple $\langle c_i, v_i\rangle$ of φ whose condition c_i is the attribute, without affecting the value of φ on $\mathtt{obs}$. After the monotone-term 1-decision list φ has been thus processed to obtain φ', all remaining attributes in φ' will be masked in $\mathtt{obs}$. By construction, $\mathtt{val}(\varphi \mid \mathtt{obs}) = \mathtt{val}(\varphi' \mid \mathtt{obs})$, and clearly $\mathtt{val}(\varphi' \mid \mathtt{obs})$ is $\mathtt{1}$ if all decisions in φ' are the value $\mathtt{1}$, is $\mathtt{0}$ if all decisions in φ' are the value $\mathtt{0}$, and is $*$ otherwise.

Condition (ii) of Theorem 5.1 follows directly from the conclusions above. For Condition (iii) of Theorem 5.1, assume that χ is satisfiable, and consider any set of literals τ over V that induces a satisfying truth-assignment for χ. Let $\varphi_\tau \triangleq \prod\{\langle a(\bar{l}), \mathtt{0}\rangle\langle a(l), \mathtt{1}\rangle \mid l \in \tau\}\langle\top, \mathtt{0}\rangle$, where multiplication between tuples is taken to correspond to their noncommutative concatenation, and l is taken to traverse τ in some fixed order over the literals over V. By construction, for every $v_i \in V$, the conditions x_i^+, x_i^- appear in φ_τ, with the condition appearing second with a corresponding value $\mathtt{1}$ being the one associated with the truth-value of v_i as determined by τ. By inspection, $\mathtt{val}(\varphi_\tau \mid \mathtt{exm}) = \mathtt{exm}[t]$ for each example $\mathtt{exm}$ in the set of examples that contains exactly the following:

(i) the example $\mathtt{exm}_{zero}$ that assigns the value $\mathtt{0}$ to attributes in $\mathcal{A}$;
(ii) for every i: $1 \leqslant i \leqslant n$, the example $\mathtt{exm}_{var(i,0)}$ that assigns the value $\mathtt{1}$ to attributes x_i^+, x_i^-, the value $\mathtt{0}$ to attribute x_t, and the value $\mathtt{0}$ to attributes in $\mathcal{A} \setminus \{x_i^+, x_i^-, x_t\}$;
(iii) for every i: $1 \leqslant i \leqslant n$, the example $\mathtt{exm}_{var(i,1)}$ that assigns the value $\mathtt{1}$ to the single attribute $x_i^\pm$ in the set $\{x_i^+, x_i^-\} \cap \{a(l) \mid l \in \tau\}$, the value $\mathtt{1}$ to attribute x_t, and the value $\mathtt{0}$ to attributes in $\mathcal{A} \setminus \{x_i^\pm, x_t\}$;
(iv) for every j: $1 \leqslant j \leqslant m$, the example $\mathtt{exm}_{cls(j)}$ that assigns the value $\mathtt{1}$ to all attributes $\mathcal{A}_{cls(j)}$ in the set $\{a(l_{j,1}), a(l_{j,2}), a(l_{j,3})\} \cap \{a(l) \mid l \in \tau\}$, the value $\mathtt{1}$ to attribute x_t, and the value $\mathtt{0}$ to attributes in $\mathcal{A} \setminus (\mathcal{A}_{cls(j)} \cup \{x_t\})$.

Define the probability distribution $\mathcal{D}$ that returns each of these examples with equal probability. Define the masking process $\mathtt{mask}$ that maps each example to the observation in $\mathcal{O}(\chi)$ with the same subscript with probability 1. Clearly, $\mathtt{mask}(\mathcal{D})$ is uniform over $\mathcal{O}(\chi)$, and the target attribute x_t is expressed by the monotone-term 1-decision list $\varphi_\tau \in \mathcal{C}$ w.r.t. the probability distribution $\mathcal{D}$; thus, $\mathcal{D}$ supports $\mathcal{C}$ for x_t, and φ_τ is the target concept for x_t under $\mathcal{D}$. Condition (iii) of Theorem 5.1 follows, and the proof is complete. □

The intractability result of Theorem 5.3 provides yet another indication that learnability from partial observations is harder than learnability from complete observations. This indication remains true even when learnability from complete observations is restricted to the use of statistical queries [15], or the use of complete observations with random classification noise [1].

6. Sensor-restricted learnability

We have taken the approach that in many domains an agent cannot a priori make any assumptions on the nature of information loss that results from its imperfect sensors. This premise is reflected in the definition of learnability that we have introduced, which asks that learning be possible for every masking process. In this section we turn our attention to domains where some bias exists on the way information is hidden when an agent senses its environment. Such a bias exists, for instance, in the way the human eye provides information on our surroundings in a spatially-dependent manner, hiding the values of those properties of the environment that lie outside the range of our sight. A cryptanalyst attempting to break some decrypting device through the use of probes, may gain some insight on the internal workings of the device by obtaining readings independently at random from each of the probes attached to the device. A piece of text, viewed as an appearance of some underlying reality, presumably hides information asymmetrically, so that, for instance, the properties of the underlying reality that are false are hidden more often than those that are true.

Bias on an agent's sensors may be captured by letting the masking process that models them be a member of a class $\mathcal{S}$ of masking processes, known as the ***sensor class***; the class contains the possible masking processes out of which one is used to obtain observations. Consistent learnability may then be redefined so that learning will be expected to be successful only if the employed masking process is a member of $\mathcal{S}$; the sensor class $\mathcal{S}$ is available to the learner, in the same way that the concept and hypothesis classes are.

Definition 6.1 *(Consistent learnability with restricted sensor class).* An algorithm $\mathcal{L}$ is a ***consistent learner for*** a learning task $\langle x_t, \mathcal{C}, \mathcal{H} \rangle$ over $\mathcal{A}$ ***with sensors in*** $\mathcal{S}$ if for every probability distribution $\mathcal{D}$ supporting $\mathcal{C}$ for x_t, every masking process $\mathtt{mask} \in \mathcal{S}$, every real number $\delta \in (0,1]$, and every real number $\varepsilon \in (0,1]$, algorithm $\mathcal{L}$ has the following property: given access to $\mathcal{A}$, $\langle x_t, \mathcal{C}, \mathcal{H} \rangle$, $\mathcal{S}$, δ, ε, and an oracle returning observations drawn from $\mathtt{mask}(\mathcal{D})$, algorithm $\mathcal{L}$ runs in time polynomial in $1/\delta$, $1/\varepsilon$, $|\mathcal{A}|$, and the size of the target concept for x_t under $\mathcal{D}$, and returns, with probability $1-\delta$, a hypothesis $h \in \mathcal{H}$ that is $(1-\varepsilon)$-consistent with x_t under $\mathcal{D}$ and $\mathtt{mask}$. The concept class $\mathcal{C}$ over $\mathcal{A} \setminus \{x_t\}$ is ***consistently learnable on*** the target attribute $x_t \in \mathcal{A}$ ***by*** the hypothesis class $\mathcal{H}$ over $\mathcal{A} \setminus \{x_t\}$ ***with sensors in*** in the sensor class $\mathcal{S}$ if there exists a consistent learner for $\langle x_t, \mathcal{C}, \mathcal{H} \rangle$ over $\mathcal{A}$ with sensors in $\mathcal{S}$.

The existence of a restricted sensor class $\mathcal{S}$ may critically affect what is learnable, and thus, what information can be recovered in partial observations. First, the restrictions obeyed by $\mathcal{S}$ may guarantee that the agent's sensors do not hide information in an entirely arbitrary manner, due to the exclusion of certain sensors from $\mathcal{S}$. This might, then, imply that the agent obtains more information than what would have been the case had $\mathcal{S}$ been unrestricted. Second, the restricted sensor class $\mathcal{S}$ may allow the agent to employ a learning algorithm tailored to $\mathcal{S}$, while no learning algorithm may be known, or even *exist*, for the general case. Through learned rules the agent may then be able to recover yet more information.

6.1. Parameterized sensor classes

It is conceivable for an agent to have a bias on the characteristics of its sensors that depends on the structure of its environment. Consider, for instance, a student in an introductory Artificial Intelligence course, during the lecture that discusses what an "agent" is. The teacher, acting as the student's sensors, presents positive and negative instances of "agents", along with various properties of the entity depicted in each instance. In trying to learn what constitutes an "agent", the student has a bias as to the sensing process through which appearances are obtained. For one, the sensing process never hides information on the property of interest that states whether an entity is an "agent" or not; this bias is readily representable in terms of a restricted sensor class, as per Definition 6.1. The student, however, has an additional, more subtle, type of bias on the sensing process: it never hides those properties that are important in defining what an "agent" is; the teacher ensures that this is the case. The bias on the type of the student's sensors *depends* on the definition of an "agent", and this bias would have been different had the definition been different.

Formalizing a ***structure-dependent sensor class*** is straightforward. We simply update Definition 6.1 so that $\mathcal{S}$ is the union $\bigcup_{c \in \mathcal{C}} \mathcal{S}_c$ of subclasses, one for each possible structure $c \in \mathcal{C}$ of the environment, and then ask that learning succeeds for every masking process $\mathtt{mask} \in \mathcal{S}_c$, where c is the target concept for x_t under $\mathcal{D}$. The learning algorithm is given access only to the sensor class $\mathcal{S}$, and not the particular subclass $\mathcal{S}_c$, since the actual structure of the agent's environment remains unknown. One may, in fact, generalize the definition even further, by allowing ***distribution-dependent sensor classes***, where the bias on the characteristics of an agent's sensors depends not only on the structure of the agent's environment, but also on the precise probability distribution from which the reality is obtained; note that this probability distribution determines also the structure of the environment. We may then update Definition 6.1 so that $\mathcal{S}$ is the union $\bigcup_{\mathcal{D}} \mathcal{S}_{\mathcal{D}}$ of subclasses, one for each possible probability distribution $\mathcal{D}$, and ask that learning succeeds for every masking process $\mathtt{mask} \in \mathcal{S}_{\mathcal{D}}$. Similar to the case of structure-dependent sensor classes, the learning algorithm is still given access only to the sensor class $\mathcal{S}$, and not the particular subclass $\mathcal{S}_{\mathcal{D}}$.

Appearances provided by teachers do not only hide information about the underlying reality, but may also convey additional information on the structure of the environment that would not normally be available. A teacher presenting an entity as a positive or negative instance of what an "agent" is, presents only a subset of the entity's properties, but also conveys the message that hidden information is irrelevant. This additional piece of information would *not* have been available to the student had complete information on the entity been presented. In a teacher-assisted learning context, the "don't know" interpretation of hidden properties does no longer characterize the nature of the value $*$. Instead, depending on the setting, the value $*$ may be better interpreted as a new, distinct, and information-baring value, indicating, for instance, a value that is "irrelevant", "the most probable", "non-deducible", "costly to obtain", or "always hidden"; in all these cases the value $*$ provides implicit information that may, in fact, play a critical role in facilitating learnability.

As a proof of concept, consider the concept class $\mathcal{C}$ of formulas in disjunctive normal form. The learnability of $\mathcal{C}$ by any hypothesis class under the PAC semantics remains one of the long-standing open questions in Computational Learning Theory, while the proper learnability of $\mathcal{C}$ when the DNF formulas are restricted to contain at most $k \in \mathbb{N}$ terms is known to be intractable for every constant $k \geqslant 2$, unless $\mathtt{RP} = \mathtt{NP}$ [22]. Yet, a teacher with the power to determine which parts of a randomly drawn example will be made visible to a student might assist the learning process by hiding information that is irrelevant in each example. More precisely, given the learning task $\langle x_t, \mathcal{C}, \mathcal{H} \rangle$ over $\mathcal{A}$, where $\mathcal{C}$ and $\mathcal{H}$ are classes of k-term DNF formulas, consider a teacher that is modelled by the following structure-dependent sensor class $\mathcal{S}_{relevant} = \bigcup_{c \in \mathcal{C}} \mathcal{S}_c$, where for every masking process $\mathtt{mask} \in \mathcal{S}_c$, and every observation $\mathtt{obs}$ in the range of $\mathtt{mask}$, the target attribute x_t is not masked in $\mathtt{obs}$, if $\mathtt{obs}[t] = 1$ then a maximal subset of attributes such that $\mathtt{val}(c \mid \mathtt{obs}) = 1$ is masked in $\mathtt{obs}$, and if $\mathtt{obs}[t] = 0$ then no attribute is masked in $\mathtt{obs}$. The next result shows the power of structure-dependent sensor classes.

Theorem 6.1 *(Proper consistent learnability of k-term DNF formulas with sensors in $\mathcal{S}_{relevant}$). The concept class $\mathcal{C}$ of formulas over $\mathcal{A} \setminus \{x_t\}$ in disjunctive normal form with at most $k \in \mathbb{N}$ terms is properly consistently learnable on the target attribute $x_t \in \mathcal{A}$ with sensors in $\mathcal{S}_{relevant}$.*

Proof. We construct an algorithm $\mathcal{L}$ as follows. Fix a probability distribution $\mathcal{D}$ supporting $\mathcal{C}$ for x_t, with c being the target concept for x_t under $\mathcal{D}$, a masking process $\mathtt{mask} \in \mathcal{S}_c \subseteq \mathcal{S}_{relevant}$, a real number $\delta \in (0,1]$, and a real number $\varepsilon \in (0,1]$. Then, algorithm $\mathcal{L}$, given access to $\mathcal{A}$, $\langle x_t, \mathcal{C}, \mathcal{C}\rangle$, $\mathcal{S}_{relevant}$, δ, ε, and an oracle returning observations drawn from $\mathtt{mask}(\mathcal{D})$, proceeds as follows:

> Algorithm $\mathcal{L}$ draws a sample $\mathcal{O}$ of a number of observations (to be determined later) from the oracle, constructs and returns the hypothesis $h = \bigvee_{\mathtt{obs} \in \mathcal{O}; \mathtt{obs}[t]=1}((\bigwedge_{x_i \in \mathcal{A}\setminus\{x_t\}; \mathtt{obs}[i]=1} x_i) \wedge (\bigwedge_{x_i \in \mathcal{A}\setminus\{x_t\}; \mathtt{obs}[i]=0} \overline{x_i}))$, and terminates.

We now prove that algorithm $\mathcal{L}$ is a consistent learner for the learning task $\langle x_t, \mathcal{C}, \mathcal{C}\rangle$ over $\mathcal{A}$ with sensors in $\mathcal{S}_{relevant}$. We show that: the returned hypothesis $h \in \mathcal{C}$; h, with probability $1-\delta$, is $(1-\varepsilon)$-consistent with x_t under $\mathcal{D}$ and $\mathtt{mask}$; and algorithm $\mathcal{L}$ runs in time polynomial in $1/\delta$, $1/\varepsilon$, $|\mathcal{A}|$, and $size(c)$.

Note first that each observation $\mathtt{obs} \leftarrow \mathtt{mask}(\mathcal{D})$ with $\mathtt{obs}[t] = 1$ encodes one of the terms of the DNF formula c. Indeed, by construction of $\mathtt{mask}$, whenever $\mathtt{obs}[t] = 1$, it also holds that $\mathtt{val}(c \mid \mathtt{obs}) = 1$, and thus at least one term of c evaluates to 1 on $\mathtt{obs}$. Thus, none of the attributes of this term are masked in $\mathtt{obs}$. By construction of $\mathtt{mask}$ the rest of the attributes are masked. It follows that $(\bigwedge_{x_i \in \mathcal{A}\setminus\{x_t\}; \mathtt{obs}[i]=1} x_i) \wedge (\bigwedge_{x_i \in \mathcal{A}\setminus\{x_t\}; \mathtt{obs}[i]=0} \overline{x_i})$ is precisely a term of c, and hence h is a DNF formula with a subset of the terms in c; in particular, $h \in \mathcal{C}$. Therefore, whenever c evaluates to 0 on an observation, so does h.

Consider now an observation $\mathtt{obs} \leftarrow \mathtt{mask}(\mathcal{D})$ w.r.t. which h has a consistency conflict with x_t, so that $\{\mathtt{val}(h \mid \mathtt{obs}), \mathtt{obs}[t]\} = \{0,1\}$. By construction of $\mathtt{mask}$, it holds that $\mathtt{obs}[t] = \mathtt{val}(c \mid \mathtt{obs})$, and hence $\{\mathtt{val}(h \mid \mathtt{obs}), \mathtt{val}(c \mid \mathtt{obs})\} = \{0,1\}$. By our preceding discussion, it holds that $\mathtt{val}(c \mid \mathtt{obs}) = 1$ and $\mathtt{val}(h \mid \mathtt{obs}) = 0$. Thus, $\mathtt{obs}$ encodes a term of c that does not appear in h, and this implies that this term was not encoded in any of the observations in $\mathcal{O}$. Consequently, if the probability of drawing such an observation is more than ε, then the probability of this observation not being part of $\mathcal{O}$ is less than $(1-\varepsilon)^{|\mathcal{O}|} \leqslant e^{-\varepsilon|\mathcal{O}|}$. When $|\mathcal{O}| = \lceil (1/\varepsilon) \cdot \ln(1/\delta) \rceil$, this probability is less than δ, as needed.

The running time of algorithm $\mathcal{L}$ is clearly linear in $1/\delta$, $1/\varepsilon$, and $|\mathcal{A}|$. This concludes the proof. □

Obtaining consistent learnability results with sensors in a *sufficiently restricted* sensor class is trivial. One need ensure only that enough attributes are masked in each drawn observation so that either the target attribute is masked, or the returned hypothesis predicts "don't know". We emphasize that the proof of Theorem 6.1 does *not* rely on this technique. All masking processes in $\mathcal{S}_{relevant}$ are 0-concealing for the target attribute x_t w.r.t. any class of formulas, so that x_t is never masked, and highly consistent hypotheses are equally highly accurate (by Theorem 2.2). One may also verify that the returned hypotheses are simultaneously highly consistent and highly complete, in that they have a consistency conflict or predict "don't know" with a total probability less than the input parameter ε. Thus, even if the returned hypotheses are forced to make $\{0,1\}$ predictions during the evaluation phase (either by changing the masking process to the identity mapping, or by assigning arbitrary values to the masked attributes), the accuracy of the returned hypotheses will not suffer. Overall, Theorem 6.1 implies a learning algorithm for obtaining hypotheses that are highly accurate even when tested on complete observations, as per the PAC semantics.

Theorem 6.1 exemplifies that attempts to examine learnability in domains where teachers choose which parts of the underlying reality students are to observe, a situation that was shown to be naturally modelled through the parameterization of sensor classes, may lead to a distortion of the meaning of missing information. Such domains essentially turn the value $*$ into a third, in addition to $\{0,1\}$, distinguished value, which may then be employed to transmit side information to an agent. This side information may then allow the agent to learn in environments where learnability would be provably impossible even with complete observations, as already illustrated. Such a treatment of missing information lies outside the scope of autodidactic learnability, as demonstrated by our choice not to permit the use of parameterized sensor classes in Definition 6.1. This problem is further dealt with in the work of Greiner et al. [10].

6.2. Special classes of sensors

The assumption that the sensor class $\mathcal{S}$ is somehow restricted underlies many previous attempts in the literature to model learnability from partial observations. Most previous work implicitly assumes the existence of a teacher in the process of learning, and as such the corresponding learning settings are naturally modelled as relying on a parameterized sensor class. The various learning models differ primarily on the type of restrictions they impose on the sensor class $\mathcal{S}$, and the type of parameterization they consider. These restrictions, in turn, explain why stronger learnability results may be obtained in other learning models, when compared to those we have obtained under the autodidactic learning semantics. We revisit some of the learning models discussed in Section 3, and summarize in Table 3 how their approach of partial observability *during the learning phase* can be viewed under the unifying prism of a restricted or parameterized sensor class.

Table 3
A unified view of various models that deal with the problem of learning from partial observations. Each model is associated with the restrictions and parameterization it imposes on the sensor class.

Learning model	Imposed restrictions and parameterization on sensor class $\mathcal{S}$
[This work]	None.
[28]	$\mathcal{S} = \bigcup_{c \in C} \mathcal{S}_c$, and each $\mathcal{S}_c$ is restricted so that: for every $\mathtt{mask} \in \mathcal{S}_c$, and every $\mathtt{obs}$ in the range of $\mathtt{mask}$, $\mathtt{obs}[t] \in \{\mathtt{0}, \mathtt{1}\}$, and $\mathtt{obs}[t] = \mathtt{1}$ if and only if $\mathtt{val}(c \mid \mathtt{obs}) = \mathtt{1}$.
[9]	$\mathcal{S} = \bigcup_{c \in C} \mathcal{S}_c$, and each $\mathcal{S}_c$ is restricted so that: for every $\mathtt{mask} \in \mathcal{S}_c$, and every $\mathtt{obs}$ in the range of $\mathtt{mask}$, $\mathtt{obs}[t] = \mathtt{val}(c \mid \mathtt{obs})$.
[29]	None a priori, but makes an implicit assumption as in [9] for learnability to be possible.
[6]	For any given probability p, $\mathcal{S} = \{\mathtt{mask}_{bernoulli(p)}\}$ so that: for every example $\mathtt{exm}$, and every $x_i \in \mathcal{A}$, $Pr[\mathtt{obs}[i] = * \mid \mathtt{obs} \leftarrow \mathtt{mask}_{bernoulli(p)}(\mathtt{exm})] = p$.
[27]	$\mathcal{S}$ is restricted so that: for every $\mathtt{mask} \in \mathcal{S}$, and every $\mathtt{obs}$ in the range of $\mathtt{mask}$, $\mathtt{obs}[t] \in \{\mathtt{0}, \mathtt{1}\}$.
[2]	For any given model parameter k, $\mathcal{S}$ is restricted so that: for every $\mathtt{mask} \in \mathcal{S}$, and every $\mathtt{obs}$ in the range of $\mathtt{mask}$, $\lvert\{x_i \mid x_i \in \mathcal{A}; \mathtt{obs}[i] \neq *\}\rvert = k$.
[1,3,13]	*Effectively*, when the set of attributes $\mathcal{A}$ is extended to $\mathcal{A}' \cup \mathcal{A}$, $\mathcal{S} = \{\mathtt{mask}_{hide}\}$, where $\mathtt{mask}_{hide}$ is the unique masking process that maps each example to the unique observation $\mathtt{obs}$ that masks $\mathtt{exm}$ and is such that $\{x_i \mid x_i \in \mathcal{A}' \cup \mathcal{A}; \mathtt{obs}[i] \neq *\} = \mathcal{A}$.

7. Outlook and future directions

We have presented the autodidactic learning model that offers a principled treatment of partial information in a PAC-like learning setting. Although it allows the use of supervised learning techniques, the autodidactic learning model does not assume the presence of an external teacher, since supervision (i.e., the label of the target attribute) is provided only to the extent that an agent's sensors do so. Within this learning model we have shown that the principle known as Occam's Razor, and the technique of reductions among learning problems are still applicable. Through reductions we have shown that monotone and read-once formulas that are PAC learnable remain learnable even if learning examples are arbitrarily incomplete. On the other hand, parities and monotone-term 1-decision lists, which are properly PAC learnable, are not properly learnable from incomplete learning examples, even if the values of only three attributes are hidden.

Numerous questions remain open: To what extent can shallowness be used to establish further learnability results? Are one-to-many reductions more beneficial than one-to-one reduction in the context of learnability? What other general techniques can be used to establish positive or negative learnability results? Can PAC learnable concept classes of formulas that are not efficiently (e.g., general 3-CNF) or locally (e.g., general 2-CNF) evaluatable on partial observations be learned, or can a general result be proven that excludes the possibility of learning such formulas? Can learnability be improved under reasonable assumptions on the masking process, without sacrificing the autonomy of learning? Does it make sense to attempt to learn the structure of the masking process, in addition to the structure of the underlying examples? Is it possible to establish the representation-independent non-learnability of some PAC-learnable concept class? Under what conditions can a certain degree of completeness be guaranteed for learned hypotheses?

Endowing learning algorithms with certain properties would significantly improve their practical applicability: One property would be to achieve running time independent of the observation size, and dependent on the number of only the non-masked attributes. It is an easy exercise to show that the Winnow algorithm [17] for learning linear thresholds can be modified to obtain this property. Another property would be the ability to exploit information in observations where the target attribute is masked; existing techniques for semi-supervised and unsupervised learning from complete information suggest that this direction is fruitful. Noise could also be dealt with. Due to the equivalent treatment of all attributes in the autodidactic learning model, it might be harder to justify the consideration of certain forms of noise considered in the literature, such as classification noise [1]. On the other hand, random noise across all attributes could be meaningfully considered [6]. The extent to which reductions preserve noise-resilience could also be investigated. Conceivably, our obtained algorithms are, to some extent, noise-resilient since they build on existing noise-resilient PAC algorithms. Noise-resilience could alternatively be established by formulating a corresponding Statistical Query model as in the case of PAC learning [15]. Finally, it would be interesting to examine whether learning is possible from examples where attributes obey more general types of correlation than that considered in this work. The role of learned rules in this setting may then change from an *explanatory* one that explains why the value of the target attribute is what it is given the values of the remaining attributes, to a *descriptive* one that simply describes what holds in examples.

We believe that the treatment of partial observability introduced herein may provide the basis for addressing certain broader issues, both in the theoretical understanding and actual implementation of systems that sense their environment via the use of imperfect sensors, for which existing solutions may be problematic or artificial. Learning rules in parallel for multiple distinct target attributes cannot be expressed in typical PAC-like supervised learning models, because the target attribute is a priori distinguished and treated differently; this is not the case in autodidactic learning. The use of learned rules for reasoning, so that their conclusions can be chained is not meaningfully supported in learning models that assume complete information. Autodidactic learning, on the other hand, naturally accommodates reasoning as the process through which some of the missing information is completed. Finally, domains where machine learning is typically employed, such as that of the autonomous acquisition of unaxiomatized or commonsense knowledge from large corpora of text [19,30], can be understood in a conceptually cleaner manner through autodidactic learning. Text can be naturally viewed as a partial depiction of some underlying and not directly accessible reality, and, then, commonsense knowledge acquisition amounts to learning to infer what holds in this reality [21]. Some of these considerations have been investigated [20].

Acknowledgements

The author is grateful to Leslie Valiant for his advice, and for valuable suggestions and remarks on this research. Useful feedback was also received from the anonymous *IJCAI* and *AIJ* reviewers.

References

[1] Dana Angluin, Philip D. Laird, Learning from noisy examples, Machine Learning 2 (4) (April 1988) 343–370.
[2] Shai Ben-David, Eli Dichterman, Learning with restricted focus of attention, Journal of Computer and System Sciences 56 (3) (April 1998) 277–298.
[3] Avrim Blum, Prasad Chalasani, Learning switching concepts, in: Proceedings of the Fifth Annual Workshop on Computational Learning Theory (COLT'92), July 1992, pp. 231–242.
[4] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, Manfred K. Warmuth, Occam's Razor, Information Processing Letters 24 (6) (April 1987) 377–380.
[5] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, Manfred K. Warmuth, Learnability and the Vapnik–Chervonenkis dimension, Journal of the ACM 36 (4) (October 1989) 929–965.
[6] Scott E. Decatur, Rosario Gennaro, On learning from noisy and incomplete examples, in: Proceedings of the Eighth Annual Conference on Computational Learning Theory (COLT'95), July 1995, pp. 353–360.
[7] Paul Fischer, Hans-Ulrich Simon, On learning ring-sum expansions, SIAM Journal on Computing 21 (1) (February 1992) 181–192.
[8] Michael R. Garey, David S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness, W.H. Freeman & Co, New York, USA, 1979.
[9] Sally A. Goldman, Stephen S. Kwek, Stephen D. Scott, Learning from examples with unspecified attribute values, Information and Computation 180 (2) (January 2003) 82–100.
[10] Russell Greiner, Adam J. Grove, Alexander Kogan, Knowing what doesn't matter: Exploiting the omission of irrelevant data, Artificial Intelligence 97 (1–2) (December 1997) 345–380.
[11] David P. Helmbold, Robert H. Sloan, Manfred K. Warmuth, Learning integer lattices, SIAM Journal on Computing 21 (2) (April 1992) 240–266.
[12] Wassily Hoeffding, Probability inequalities for sums of bounded random variables, Journal of the American Statistical Association 58 (301) (March 1963) 13–30.
[13] Michael J. Kearns, Robert E. Schapire, Efficient distribution-free learning of probabilistic concepts, Journal of Computer and System Sciences 48 (3) (June 1994) 464–497.
[14] Michael J. Kearns, Umesh V. Vazirani, An Introduction to Computational Learning Theory, The MIT Press, Cambridge, Massachusetts, USA, 1994.
[15] Michael J. Kearns, Efficient noise-tolerant learning from statistical queries, Journal of the ACM 45 (6) (November 1998) 983–1006.
[16] Roderick J.A. Little, Donald B. Rubin, Statistical Analysis with Missing Data, 2nd ed., John Wiley & Sons, Inc., New York, USA, 2002.
[17] Nick Littlestone, Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm, Machine Learning 2 (4) (April 1988) 285–318.
[18] John McCarthy, Appearance and reality, John McCarthy's home page http://www-formal.stanford.edu/jmc/appearance.html, 30 August 2006.
[19] Loizos Michael, Leslie G. Valiant, A first experimental demonstration of massive knowledge infusion, in: Proceedings of the Eleventh International Conference on Principles of Knowledge Representation and Reasoning (KR'08), September 2008, pp. 378–388.
[20] Loizos Michael, Autodidactic learning and reasoning, PhD thesis, School of Engineering and Applied Sciences, Harvard University, USA, May 2008.
[21] Loizos Michael, Reading between the lines, in: Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI'09), July 2009, pp. 1525–1530.
[22] Leonard Pitt, Leslie G. Valiant, Computational limitations on learning from examples, Journal of the ACM 35 (4) (October 1988) 965–984.
[23] Leonard Pitt, Manfred K. Warmuth, Prediction-preserving reducibility, Journal of Computer and System Sciences 41 (3) (December 1990) 430–467.
[24] Ronald L. Rivest, Robert Sloan, A formal model of hierarchical concept learning, Information and Computation 114 (1) (1994) 88–114.
[25] Ronald L. Rivest, Learning decision lists, Machine Learning 2 (3) (November 1987) 229–246.
[26] Joseph L. Schafer, John W. Graham, Missing data: Our view of the state of the art, Psychological Methods 7 (2) (June 2002) 147–177.
[27] Dale Schuurmans, Russell Greiner, Learning default concepts, in: Proceedings of the Tenth Canadian Conference on Artificial Intelligence (AI'94), May 1994, pp. 99–106.
[28] Leslie G. Valiant, A theory of the learnable, Communications of the ACM 27 (11) (November 1984) 1134–1142.
[29] Leslie G. Valiant, Robust logics, Artificial Intelligence 117 (2) (March 2000) 231–253.
[30] Leslie G. Valiant, Knowledge infusion, in: Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI'06), July 2006, pp. 1546–1551.
[31] Leslie G. Valiant, Evolvability, Journal of the ACM 56 (1) (January 2009) 3.1–3.21.