Learning from Partial Observations*

Loizos Michael
Division of Engineering and Applied Sciences
Harvard University, Cambridge, MA 02138, U.S.A.
loizos@eecs.harvard.edu

Abstract

We present a general machine learning framework for modelling the phenomenon of missing information in data. We propose a masking process model to capture the stochastic nature of information loss. Learning in this context is employed as a means to recover as much of the missing information as is recoverable. We extend the *Probably Approximately Correct* semantics to the case of learning from partial observations with *arbitrarily* hidden attributes. We establish that simply requiring learned hypotheses to be consistent with observed values suffices to guarantee that hidden values are recoverable to a certain accuracy; we also show that, in some sense, this is an optimal strategy for achieving accurate recovery. We then establish that a number of natural concept classes, including all the classes of monotone formulas that are PAC learnable by monotone formulas, and the classes of conjunctions, disjunctions, k-CNF, k-DNF, and linear thresholds, are consistently learnable from partial observations. We finally show that the concept classes of parities and monotone term 1-decision lists are not *properly* consistently learnable from partial observations, if $\mathtt{RP} \neq \mathtt{NP}$. This implies a separation of what is consistently learnable from partial observations versus what is learnable in the complete or noisy setting.

1 Introduction

Consider the task of predicting missing entries in a medical database, given the information that is already available. How does one go about making such predictions, and what kind of guarantees might one provide on the accuracy of these predictions? The problem with which one is faced here is that of missing information in data, an arguably universal and multidisciplinary problem. Standard statistical techniques [Schafer and Graham, 2002] fail to provide a formal treatment for the general case of this problem, where the fact that information is missing might be arbitrarily correlated with the actual value of the missing information (e.g., patients exhibiting a certain symptom might be less inclined to disclose this fact).

*This work was supported by grant NSF-CCF-04-27129.

In this work we employ learning as a means of identifying structure in a domain of interest, given access to certain observations. We subsequently utilize such identified structure to recover missing information in new observations coming from the same domain. Note that the manner in which acquired knowledge may be utilized to draw conclusions is not necessarily a single step process (see, e.g., [Valiant, 2000]). Nonetheless, we focus here on examining whether even individual learned rules can be meaningfully applied on partial observations, given that such rules are *learned from observations that are partial themselves*. Studying how multiple learned rules can be chained and reasoned with to draw richer conclusions presents further challenges, and necessitates a solution to the more fundamental problem examined herein.

We present a general machine learning framework within which the problem of dealing with missing information can be understood. We formulate the notion of *masked* attributes, whose values in learning examples (e.g., patient records) are not made known to an agent. Such masked attributes account for missing information both in the given target that the agent attempts to learn (e.g., the presence of a particular disease), as well as in the learning features over which the agent's hypotheses are formed (e.g., various pieces of information from a patient's medical history). Masked attributes are determined by an *arbitrary* stochastic process that induces for each example a possibly different but fixed distribution over partial observations to which the example is mapped (see also [Schuurmans and Greiner, 1994]); this is intended to capture situations such as the probabilistic failure or inability of an agent's sensors to provide readings. We extend the *Probably Approximately Correct* learning semantics [Valiant, 1984] to apply to the described situation. A salient feature of the extension we propose is the lack of need for specially prepared learning materials; the agent simply utilizes whatever information is made available through the masking process. We call this type of learning *autodidactic* to emphasize that although the agent might still employ supervised learning techniques, this is done without the presence of a teacher explicitly providing the agent with "labelled instances" during the learning phase.

We propose *consistency* as an intuitive measure of success of the learning process. An agent faced with partial observations needs to produce hypotheses that do not contradict what is actually observed; the values of masked attributes need not be predicted correctly. In addition, hypotheses that do not as-

sume a definite value due to masked attributes need not make a prediction. We allow, thus, the possibility of "don't know" predictions, but restrict such predictions in a natural manner, providing a notion of *completeness* of the prediction process.

Following the presentation of our framework, we discuss *accuracy* as an alternative measure of success, whereby an agent is expected to correctly predict the values of masked attributes. We show that the success of an agent in this stricter setting might be completely impaired, depending on how *concealing* the masking process is (i.e., how adversarially information is hidden from the agent). On the positive side, we show that to the degree allowed by the masking process, an agent can perform optimally in making accurate predictions, by simply making consistent predictions. This surprising relation between the two measures of success allows an agent to focus on the more natural task of learning consistently, while not losing anything with respect to predicting accurately.

We then examine consistent learnability more closely. We define a notion of reduction between learning tasks, and establish that any concept class of monotone formulas that is PAC learnable by some hypothesis class of monotone formulas is also consistently learnable from partial observations; the result is obtained by reducing the learning task to one over complete observations. Through a second reduction we show that the concept classes of conjunctions, disjunctions, k-CNF, k-DNF, and linear thresholds over literals, are all *properly* consistently learnable from partial observations.

On the negative side, we show that the set of consistently learnable concept classes is a subset of the PAC learnable concept classes. We continue to prove that the concept classes of parities and monotone term 1-decision lists are not *properly* consistently learnable from partial observations, given that the widely held complexity assumption $\mathtt{RP} \neq \mathtt{NP}$ is true. The intractability of properly learning monotone term 1-decision lists from partial observations provides a partial answer to a question posed by Rivest [1987]. Our intractability results establish separations between our model of consistent learnability from partial observations, and the existing models of PAC learnability [Valiant, 1984] and learnability in the presence of random classification noise [Angluin and Laird, 1987].

We assume the reader is familiar with basic PAC learning terminology (see, e.g., [Kearns and Vazirani, 1994]). Proofs are only briefly discussed in this paper due to lack of space.

2 The Learning Framework

In the PAC learning model [Valiant, 1984], a set of boolean variables $\{x_1, x_2, \ldots, x_n\}$ represents the attributes of the environment. A concept c is a boolean formula over the boolean variables. An example for the concept c is a truth-assignment to the boolean variables, drawn from an underlying probability distribution $\mathcal{D}$, paired with the induced truth-value of c.

Such a treatment distinguishes the target attribute from the attributes acting as learning features for that target. As a more natural and better suited approach for *autodidactic learning*, where target attributes are not externally "labelled", we consider examples that treat all attributes equally as properties of the environment. The attribute acting as a learning target need only be defined as part of the learning task one undertakes.

Definition 2.1 (Examples and Observations) *Consider any non-empty finite set $\mathcal{A}$ of **attributes**. An **example** over $\mathcal{A}$ is a vector $\mathtt{exm} \in \{0,1\}^{|\mathcal{A}|}$. An **observation** over $\mathcal{A}$ is a vector $\mathtt{obs} \in \{0,1,*\}^{|\mathcal{A}|}$. An observation $\mathtt{obs}$ **masks** an example $\mathtt{exm}$ if $\mathtt{obs}[i] \in \{\mathtt{exm}[i], *\}$ for every attribute $x_i \in \mathcal{A}$. An attribute $x_i \in \mathcal{A}$ is **masked** in an observation $\mathtt{obs}$ if $\mathtt{obs}[i] = *$. A **masking process** is a (stochastic) function $\mathtt{mask} : \{0,1\}^{|\mathcal{A}|} \to \{0,1,*\}^{|\mathcal{A}|}$ that maps each example $\mathtt{exm}$ to some observation $\mathtt{obs}$ that masks $\mathtt{exm}$.*

Examples define the "truth" about the environment. Such examples are drawn from some underlying fixed probability distribution $\mathcal{D}$ that is unknown to the agent. Unlike standard PAC learning, the agent does not directly observe such examples, but only masked versions of the examples. We denote by $\mathtt{mask}(\mathcal{D})$ the induced distribution over these observations.

The stochastic masking process can be understood in two ways. If attributes correspond to an agent's sensors, masking corresponds to the stochastic failure of these sensors to provide readings. If attributes correspond to properties of the environment, masking corresponds to an agent's inability to simultaneously sense all properties. In either case, masking induces for each example $\mathtt{exm}$ a possibly different but fixed distribution $\mathtt{mask}(\mathtt{exm})$ over observations; the induced distributions remain unknown to the agent, and so does $\mathtt{mask}(\mathcal{D})$.

Masked attributes have their values hidden, without any connotations. In particular, a masked attribute is not to be understood as "non-deducible" from the rest of the attributes. The goal of an agent is not to deduce that a masked attribute is assigned the value $*$, but rather to deduce the truth-value of the masked attribute according to the underlying masked example. This is a rather non-trivial, and sometimes impossible, task, depending on the masking process being considered.

Definition 2.2 (Formulas) *A **formula** $f(x_{i_1}, \ldots, x_{i_k})$ over $\mathcal{A}$ is a function $f : \{0,1\}^k \to \{0,1\}$ whose arguments are associated with the attributes $x_{i_1}, \ldots, x_{i_k} \in \mathcal{A}$. The **value** of $f(x_{i_1}, \ldots, x_{i_k})$ given an example $\mathtt{exm}$ is defined to be $\mathtt{val}(f(x_{i_1}, \ldots, x_{i_k}) \,|\, \mathtt{exm}) \triangleq f(\mathtt{exm}[i_1], \ldots, \mathtt{exm}[i_k])$. The **value** of $f(x_{i_1}, \ldots, x_{i_k})$ given an observation $\mathtt{obs}$, denoted by $\mathtt{val}(f(x_{i_1}, \ldots, x_{i_k}) \,|\, \mathtt{obs})$, is defined to be the common value of the formula given all examples masked by $\mathtt{obs}$, in case such a common value exists, or $*$ otherwise.*

An agent's task of identifying structure in its environment can be made precise as the problem of learning how a certain *target attribute* in $\mathcal{A}$ can be expressed as a formula over other attributes in $\mathcal{A}$, as these are perceived through the agent's sensors.[1] To study learnability, one usually assumes that the target attribute is indeed expressible as such a formula, called the *target concept*, and that the target attribute always assumes a truth-value according to the target concept. The described setting is captured by the following definitions.

Definition 2.3 (Formula Equivalence) *Formulas φ_1 and φ_2 over $\mathcal{A}$ are **equivalent** w.r.t. a probability distribution $\mathcal{D}$ if $Pr\left(\mathtt{val}(\varphi_1 \,|\, \mathtt{exm}) = \mathtt{val}(\varphi_2 \,|\, \mathtt{exm}) \mid \mathtt{exm} \leftarrow \mathcal{D}\right) = 1$.*

[1] More generally, one can consider how a formula over attributes can be expressed as a formula over other attributes. The approach is similar, and our definitions and results apply largely unchanged.

Definition 2.4 (Supported Concepts) *A **concept class** over $\mathcal{A}$ is a set $\mathcal{C}$ of formulas over $\mathcal{A}$. A probability distribution $\mathcal{D}$ **supports** $\mathcal{C}$ for an attribute x_t if x_t is equivalent to some formula $c \in \mathcal{C}$ w.r.t. $\mathcal{D}$; c is the **target concept** for x_t under $\mathcal{D}$.*

Supported concept classes essentially encode a known or assumed bias on the probability distribution from which examples are drawn. This imposes constraints on the examples, in what is perhaps the simplest possible manner that still facilitates learnability. Assuming such a bias, the goal of an agent is then to identify a formula from some hypothesis class, that is consistent with the target attribute with high probability.

Definition 2.5 (Learning Tasks) *A **learning task** over $\mathcal{A}$ is a triple $\langle x_t, \mathcal{C}, \mathcal{H} \rangle$, where x_t is an attribute in $\mathcal{A}$, $\mathcal{C}$ is a concept class over $\mathcal{A}$, and $\mathcal{H}$ a hypothesis class of formulas over $\mathcal{A}$.*

We omit writing the set of attributes $\mathcal{A}$ over which a learning task is defined, when this does not introduce ambiguities.

Definition 2.6 ($(1-\varepsilon)$-consistency) *A hypothesis h **conflicts** with a target attribute $x_t \in \mathcal{A}$ w.r.t. an observation $\mathtt{obs}$ if $\{\mathtt{val}(h \mid \mathtt{obs}), \mathtt{val}(x_t \mid \mathtt{obs})\} = \{0, 1\}$. A hypothesis h is $(1-\varepsilon)$**-consistent** with a target attribute $x_t \in \mathcal{A}$ under a probability distribution $\mathcal{D}$ and a masking process $\mathtt{mask}$ if*

$$Pr\left(\{\mathtt{val}(h \mid \mathtt{obs}), \mathtt{val}(x_t \mid \mathtt{obs})\} = \{0, 1\} \mid \mathtt{exm} \leftarrow \mathcal{D}; \mathtt{obs} \leftarrow \mathtt{mask}(\mathtt{exm})\right) \leq \varepsilon.$$

Recall that formulas might evaluate to $*$ given an observation. We interpret this value as a "don't know" prediction, and such a prediction is always consistent with a target attribute. Similarly, a value of $*$ for an attribute is interpreted as a "don't know" sensor reading, and every prediction is consistent with such a sensor reading. That is to say, as long as the prediction coming through a hypothesis and the sensor reading do not directly conflict by producing different $\{0, 1\}$ values, there is no inconsistency *at the observational level*.

It is important to note that the ability to make "don't know" predictions cannot be abused by an agent. Every hypothesis is necessarily a formula, which assumes a definite $\{0, 1\}$ value whenever sufficiently many of its arguments are specified. It only evaluates to $*$ given an observation, when its value on the actual underlying example that was masked to obtain the observation cannot be determined. Thus, our framework accounts for an implicit notion of *completeness*, by imposing a natural restriction on the "don't know" predictions.

Definition 2.7 (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 $\mathtt{mask}$, every real number $\delta : 0 < \delta \leq 1$, and every real number $\varepsilon : 0 < \varepsilon \leq 1$, the algorithm 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 with probability $1-\delta$ returns a hypothesis $h \in \mathcal{H}$ that is $(1-\varepsilon)$-consistent with x_t under $\mathtt{mask}(\mathcal{D})$. The concept class $\mathcal{C}$ over $\mathcal{A}$ is **consistently learnable** on x_t by $\mathcal{H}$ if there exists a consistent learner for $\langle x_t, \mathcal{C}, \mathcal{H} \rangle$ over $\mathcal{A}$.*

3 Consistent Learners vs. Accurate Predictors

We have taken the approach that learned hypotheses are expected to be *consistent* with observations, a natural generalization of the respective requirements of PAC learning. Such hypotheses will correctly predict the values of non-masked attributes that are artificially (for the purposes of analysis) "obscured" in an observation, after the observation is drawn. In some sense this is the best one can hope for. If an agent never gets to observe parts of its environment, then it can only form hypotheses that in the best case are consistent with its observations, although they might not agree with the underlying masked examples. This is reminiscent of developing physical theories by finding laws that are consistent with what we observe, without this implying that our current, past, or future physical theories are actually the "correct" ones. Hypotheses developed in this manner are, of course, used to make predictions on masked attributes of the world. Humans go into great lengths to subsequently obtain the values of such masked attributes, so as to experimentally validate a physical theory.

In the context of this work we study whether developed theories, or hypotheses, that are consistent with the partial observations of an agent, would actually make accurate predictions on a hypothetical validation experiment. That is, given an observation $\mathtt{obs}$ masking an example $\mathtt{exm}$, and an attribute x_t that is masked in $\mathtt{obs}$, we wish to examine whether it is possible to predict $\mathtt{exm}[t]$, and thus *accurately* (and not simply consistently) "fill-in" the missing information in $\mathtt{obs}$.

Definition 3.1 ($(1-\varepsilon)$-accuracy) *A hypothesis h 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 $\mathtt{mask}$ if*

$$Pr\left(\{\mathtt{val}(h \mid \mathtt{obs}), \mathtt{val}(x_t \mid \mathtt{exm})\} = \{0, 1\} \mid \mathtt{exm} \leftarrow \mathcal{D}; \mathtt{obs} \leftarrow \mathtt{mask}(\mathtt{exm})\right) \leq \varepsilon.$$

Hypotheses might still evaluate to $*$ given an observation. Thus, the accuracy requirement amounts to asking that *whenever* a hypothesis predicts a $\{0, 1\}$ value, the value should be in accordance with the *actual* (rather than the observed) value of the target attribute. Identifying the conditions under which one can form accurate hypotheses is essential, in that an agent's actions yield utility based not on what the agent observes, but based on what actually holds in the agent's environment. The more informed the agent is about the actual state of its environment (either through observations or accurate predictions), the better decisions the agent might reach.

Clearly, predictions that are accurate are necessarily consistent (since it holds that $\mathtt{obs}[t] \in \{\mathtt{exm}[t], *\}$). The other direction, however, does not hold in general. Indeed, predictions on masked target attributes are always consistent, while there is no evident reason why they should also be accurate.

Theorem 3.1 (Indistinguishability in Adversarial Settings) *Consider a target attribute x_t, and a concept class $\mathcal{C}$ over $\mathcal{A} \setminus \{x_t\}$, and let $\varphi_1, \varphi_2 \in \mathcal{C}$ be such that $\varphi_1 \neq \varphi_2$. There exist probability distributions $\mathcal{D}_1, \mathcal{D}_2$ such that: (i) φ_1, φ_2 are equivalent w.r.t. neither $\mathcal{D}_1$ nor $\mathcal{D}_2$, (ii) φ_1, x_t are equivalent w.r.t. $\mathcal{D}_1$, and (iii) φ_2, x_t are equivalent w.r.t. $\mathcal{D}_2$. There also exists a masking process $\mathtt{mask}$ such that $\mathtt{mask}(\mathcal{D}_1) = \mathtt{mask}(\mathcal{D}_2)$, and no attribute in $\mathcal{A} \setminus \{x_t\}$ is masked in any drawn observation.*

Theorem 3.1 shows that examples might be masked in such a way so that two non-equivalent concepts are indistinguishable given a set of observations. In fact, it suffices to only mask the target attribute in a few (but adversarially selected)

cases for the result to go through. The non-masked attributes are also adversarially selected so that observations will imply a $\{0,1\}$ value for all formulas over $\mathcal{A} \setminus \{x_t\}$, excluding the possibility of a "don't know" prediction. Clearly, an agent has no means of identifying which of the probability distributions $\mathcal{D}_1, \mathcal{D}_2$ examples are drawn from, or equivalently, which of the formulas φ_1, φ_2 is the target concept for x_t. Thus, it is impossible for the agent to confidently return a hypothesis that is highly accurate w.r.t. x_t under $\texttt{mask}(\mathcal{D}_1) = \texttt{mask}(\mathcal{D}_2)$; either confidence or accuracy is necessarily compromised.

We note that the indistinguishability result imposes very mild restrictions on the probability distributions $\mathcal{D}_1, \mathcal{D}_2$, and the concept class $\mathcal{C}$, which implies that an adversarial choice of the masking process mask can "almost always" prove disastrous for an algorithm attempting to make accurate predictions, even if the algorithm is *computationally unbounded*, the known bias on the probability distribution is *as strong as possible* (i.e., the concept class is of cardinality two), and the hypothesis class comprises of *all formulas over* $\mathcal{A} \setminus \{x_t\}$.

The established impossibility result suggests that having an infrequently masked target attribute does not suffice to learn accurately; it is important to have an infrequently masked target attribute *in the right context*. We formalize this next.

Definition 3.2 ($(1-\eta)$-concealment) *A masking process* mask *is* $(1-\eta)$***-concealing*** *for a learning task* $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ *if* η *is the maximum value such that for every example* exm*, and every hypothesis* $h \in \mathcal{H}$

$$Pr\,(\texttt{val}(x_t \,|\, \texttt{obs}) \neq * \mid \texttt{obs} \leftarrow \texttt{mask}(\texttt{exm}); \\ \{\texttt{val}(h \,|\, \texttt{obs}), \texttt{val}(x_t \,|\, \texttt{exm})\} = \{0,1\}) \geq \eta.$$

Roughly speaking, Definition 3.2 asks that whenever a hypothesis is inaccurate, the agent will observe evidence of this fact with some probability. This generalizes the case of PAC learning, where an inaccurate hypothesis is always observed to conflict with the target attribute (which is never masked). We note that the masking process mask whose existence is guaranteed by Theorem 3.1 is necessarily 1-concealing for *every* learning task $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ with a non-trivial concept class.

Theorem 3.2 (The Relation of Consistency and Accuracy) *Consider a learning task* $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$*, and a masking process* mask *that is* $(1-\eta)$*-concealing for* $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$*. Then, (i) for every probability distribution* $\mathcal{D}$ *and hypothesis* $h \in \mathcal{H}$*,* h *is* $(1-\varepsilon/\eta)$*-accurate w.r.t.* x_t *under* $\texttt{mask}(\mathcal{D})$ *if* h *is* $(1-\varepsilon)$*-consistent with* x_t *under* $\texttt{mask}(\mathcal{D})$*, and (ii) there exists a probability distribution* $\mathcal{D}_0$ *and a hypothesis* $h_0 \in \mathcal{H}$ *such that* h_0 *is* $(1-\varepsilon/\eta)$*-accurate w.r.t.* x_t *under* $\texttt{mask}(\mathcal{D}_0)$ *only if* h_0 *is* $(1-\varepsilon)$*-consistent with* x_t *under* $\texttt{mask}(\mathcal{D}_0)$*.*

Assuming that our physical world does not adversarially hide information from us, one can interpret the above result as a partial explanation of how it is possible for humans to learn rules, and construct physical theories, that make accurate predictions in situations where nothing is known, despite the fact that learning takes place and is evaluated mostly on observations with inherently missing information.

Similarly to the case of constructing learners for noisy examples [Kearns, 1998], we assume that an algorithm is given a bound on the concealment degree of the masking process and allowed time that depends on this bound during learning.

Definition 3.3 (Accurate Predictability) *An algorithm* $\mathcal{L}$ *is an* ***accurate predictor*** *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 real number* $\eta : 0 < \eta \leq 1$*, every masking process* mask *that is* $(1-\eta)$*-concealing for* $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$*, every real number* $\delta : 0 < \delta \leq 1$*, and every real number* $\varepsilon : 0 < \varepsilon \leq 1$*, the algorithm runs in time polynomial in* $1/\eta$*,* $1/\delta$*,* $1/\varepsilon$*,* $|\mathcal{A}|$*, and the size of the target concept for* x_t *under* $\mathcal{D}$*, and with probability* $1-\delta$ *returns a hypothesis* $h \in \mathcal{H}$ *that is* $(1-\varepsilon)$*-accurate w.r.t.* x_t *under* $\texttt{mask}(\mathcal{D})$*. The concept class* $\mathcal{C}$ *over* $\mathcal{A}$ *is* ***accurately predictable*** *on* x_t *by* $\mathcal{H}$ *if there exists an accurate predictor for* $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ *over* $\mathcal{A}$*.*

It is now straightforward to show the following.

Theorem 3.3 (Consistent Learners / Accurate Predictors) *Consider a learning task* $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$*, and a masking process* mask *that is* $(1-\eta)$*-concealing for* $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$*. If algorithm* $\mathcal{L}$ *is a consistent learner for* $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$*, then algorithm* $\mathcal{L}$ *given* η *as extra input and allowed running time that grows polynomially in* $1/\eta$*, is an accurate predictor for* $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$*.*

We have thus established not only that consistent learning implies accurate predicting, but that the *same* algorithm can be used, with the only provision that the algorithm will be allowed more running time to achieve the same precision as determined by ε. The running time dependence on η can be eliminated if the following are true: *(i)* the consistent learner is such that it only uses the observations that do not mask the target attribute, and *(ii)* the induced predictor has access to an oracle that returns observations from distribution $\texttt{mask}(\mathcal{D})$, conditioned, however, on the observations not masking the target attribute. The use of such an oracle exemplifies the fact that a predictor does not require more computation to produce an accurate hypothesis, but rather more observations in order to obtain enough "labelled instances" of the target concept.

A rather intriguing implication of our results is that a consistent learner is, *without* any knowledge of the concealment degree of the masking process, also able to predict accurately, albeit with a "discounted" accuracy factor. In fact, as condition *(ii)* of Theorem 3.2 suggests, *a consistent learner is, in some sense, as accurate a predictor as possible.* Given this result, it suffices to restrict our attention to consistent learning for the rest of our study on learning from partial observations.

4 Consistently Learnable Concept Classes

The stronger learnability requirements we impose compared to PAC learning do not render learnability impossible. It is an easy exercise to show that the typical algorithm for PAC learning conjunctions [Valiant, 1984] and its analysis can be applied essentially unmodified on partial observations.

Theorem 4.1 *The concept class* $\mathcal{C}$ *of conjunctions of literals over* $\mathcal{A} \setminus \{x_t\}$ *is properly consistently learnable on* x_t*.*

4.1 One-To-Many Reductions

Reductions between learning tasks are often used to establish that certain concept classes are or are not learnable. Standard reductions map examples from one learning task to examples of a different learning task. In our case such reductions map, in general, partial observations to partial observations.

Definition 4.1 (Reductions) *The learning task $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ over $\mathcal{A}$ is **reducible** to the set $\{\langle x_t^j, \mathcal{C}^j, \mathcal{H}^j\rangle$ over $\mathcal{A}^j\}_{j=0}^{r-1}$ of learning tasks, where $r \in \mathbb{N}$ is polynomially-bounded by $|\mathcal{A}|$, if there exists an efficiently computable **hypothesis mapping** $g : \mathcal{H}^0 \times \ldots \times \mathcal{H}^{r-1} \to \mathcal{H}$, and an efficiently computable **instance mapping** $f^j : \{0,1,*\}^{|\mathcal{A}|} \to \{0,1,*\}^{|\mathcal{A}^j|}$ for every $j \in \{0,\ldots,r-1\}$, such that the following conditions hold:*

(i) for every tuple $\overline{h} \in \mathcal{H}^0 \times \ldots \times \mathcal{H}^{r-1}$ and every observation $\mathtt{obs} \in \{0,1,*\}^{|\mathcal{A}|}$, it holds that $g(\overline{h})$ conflicts with x_t w.r.t. $\mathtt{obs}$ only if there exists $j \in \{0,\ldots,r-1\}$ such that $h^j = \overline{h}[j]$ conflicts with x_t^j w.r.t. $f^j(\mathtt{obs})$;

(ii) the probability distribution $\mathtt{mask}(\mathcal{D})$ from which $\mathtt{obs}$ is drawn is such that $\mathcal{D}$ supports $\mathcal{C}$ for x_t only if for every $j \in \{0,\ldots,r-1\}$ there exists an induced probability distribution $\mathtt{mask}^j(\mathcal{D}^j)$ from which $f^j(\mathtt{obs})$ is drawn such that $\mathcal{D}^j$ supports $\mathcal{C}^j$ for x_t^j, and the size of the target concept for x_t^j under $\mathcal{D}^j$ is polynomially-bounded by $|\mathcal{A}|$ and the size of the target concept for x_t under $\mathcal{D}$.

Roughly, the two conditions guarantee that *(i)* learned hypotheses can be meaningfully employed in the original task, and that *(ii)* observations in the resulting tasks can be obtained by masking examples drawn from appropriate distributions.

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 $\mathcal{A}^j\}_{j=0}^{r-1}$. The concept class $\mathcal{C}$ is consistently learnable on x_t by $\mathcal{H}$ if for every $j \in \{0,\ldots,r-1\}$, the concept class $\mathcal{C}^j$ is consistently learnable on x_t^j by $\mathcal{H}^j$.*

The following special case is of particular interest, in that observations in the resulting learning tasks are complete.

Definition 4.2 (Total Reductions) *A reduction is **total** if for every $j \in \{0,\ldots,r-1\}$, $f^j : \{0,1,*\}^{|\mathcal{A}|} \to \{0,1\}^{|\mathcal{A}^j|}$.*

4.2 Shallow-Monotone Formulas

We establish a reduction between certain classes of formulas.

Definition 4.3 (Shallow-Monotonicity) *A formula φ is **shallow-monotone** w.r.t. a set $\mathcal{M}$ of **substitutions** if the process of substituting an attribute $x'_{i(\psi)}$ for every sub-formula ψ of φ such that $x'_{i(\psi)}/\psi \in \mathcal{M}$, produces a monotone formula; denote by $\mathtt{basis}(\varphi \mid \mathcal{M})$ the resulting (monotone) formula. A set $\mathcal{F}$ of formulas is **shallow-monotone** w.r.t. a set $\mathcal{M}$ of substitutions if every formula $\varphi \in \mathcal{F}$ is shallow-monotone w.r.t. $\mathcal{M}$; we define $\mathtt{basis}(\mathcal{F} \mid \mathcal{M}) \triangleq \{\mathtt{basis}(\varphi \mid \mathcal{M}) \mid \varphi \in \mathcal{F}\}$.*

We implicitly assume that the substitution process replaces distinct new attributes for distinct sub-formulas of φ, and that the resulting formula is entirely over these new attributes.

Clearly, every set $\mathcal{F}$ of formulas is shallow-monotone w.r.t. some set $\mathcal{M}$ of substitutions. The emphasis of Definition 4.3 is on the choice of $\mathcal{M}$, and the corresponding basis of $\mathcal{F}$ w.r.t. $\mathcal{M}$. Note, for instance, that the class of k-CNF formulas for some constant $k \in \mathbb{N}$ has a basis that comprises of conjunctions, and this basis is w.r.t. a set of substitutions that is only polynomially large in the number of attributes over which the k-CNF formulas are defined (since exactly one substitution is required for each of the polynomially many possible clauses).

Theorem 4.3 (Reduction to Monotone Classes) *The learning task $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ over $\mathcal{A}$ is reducible to the 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) $\mathcal{M}$ is computable in time polynomial in $|\mathcal{A}|$, (ii) every sub-formula substituted under $\mathcal{M}$ can be evaluated given an observation in time polynomial in $|\mathcal{A}|$, and (iii) $x'_t = \mathtt{basis}(x_t \mid \mathcal{M})$, $\mathcal{C}' = \mathtt{basis}(\mathcal{C} \mid \mathcal{M})$, and $\mathcal{H}' = \mathtt{basis}(\mathcal{H} \mid \mathcal{M})$.*

An immediate corollary of Theorems 4.1 and 4.3 is that the concept class of k-CNF formulas is properly consistently learnable for every constant $k \in \mathbb{N}$.

4.3 Learning Monotone Formulas

We employ reductions to establish certain learnability results.

Theorem 4.4 (Total Self-Reduction) *The learning task $\langle x_t, \mathcal{C}, \mathcal{H}\rangle$ over $\mathcal{A}$ is total reducible to the 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 $g(\cdot)$ is the identity mapping, if $\mathcal{C}$ and $\mathcal{H}$ are classes of monotone formulas over $\mathcal{A} \setminus \{x_t\}$, and $\top \notin \mathcal{C}$.*[2]

Proof Idea: By monotonicity, any formula in $\{x_t\} \cup \mathcal{C} \cup \mathcal{H}$ that assumes a $\{0,1\}$ value given $\mathtt{obs}$ retains its value when masked attributes are mapped to $\mathtt{val}(x_t \mid \mathtt{obs}) \in \{0,1\}$. □

Theorem 4.4 establishes a rather surprising fact for monotone formulas: consistently learning from partial observations reduces to consistently learning the *same* concept class from complete observations. Equally intriguing is the fact that hypotheses learned (from complete observations) for the resulting task, apply *unmodified* for making predictions (on partial observations) in the original task. The preceding facts nicely complement Theorem 3.2, which establishes that consistently learned hypotheses are also as accurate as possible. Together, Theorems 3.2 and 4.4 imply that a concrete strategy to predict accurately on partial observations is to simply assign appropriate default truth-values to masked attributes, consistently learn from the resulting complete observations, and then employ the learned hypothesis unchanged to make predictions.

A technical point worth discussing here is the encoding of the value of the target attribute in certain attributes of the resulting task. Observe that an agent learning in the resulting task, although agnostic to this fact, uses the "label" of the example in a much more involved manner than its standard use as a means to test the predictions of a hypothesis. What makes the result established by the reduction non-trivial, is the fact that the hypothesis does not depend on the target attribute in the context of the original task. In some sense, we allow an agent to use an example's "label" in an involved manner when learning, but require that the hypothesis eventually employed for making predictions does not depend on the target attribute.

The next corollary follows from Theorems 4.3 and 4.4, and the PAC learnability of certain monotone concept classes.

Corollary 4.5 *Each concept class $\mathcal{C} \in \{$conjunctions, disjunctions, k-CNF, k-DNF, linear thresholds$\}$ of literals over $\mathcal{A} \setminus \{x_t\}$ is properly consistently learnable on x_t.*

[2] Assuming $\top \notin \mathcal{C}$ is without loss of generality, since sampling can be used to determine w.h.p. whether the target concept is a tautology, and the reduction can be used only when this is not the case.

The result holds, more generally, for any class of monotone formulas PAC learnable by *some* class of monotone formulas.

5 Negative Results in Consistent Learning

Consistent learnability is at least as strong as PAC learnability, as it requires learning under arbitrary masking processes, including the trivial identity masking process. This observation provides a basic upper bound on consistent learnability.

Theorem 5.1 *A concept class $\mathcal{C}$ over $\mathcal{A} \setminus \{x_t\}$ is consistently learnable on the target attribute x_t by a hypothesis class $\mathcal{H}$ over $\mathcal{A} \setminus \{x_t\}$ only if $\mathcal{C}$ is PAC learnable by $\mathcal{H}$.*

Theorem 5.1 implies that any concept class known not to be PAC learnable (possibly under some assumptions), is also not consistently learnable (under the same assumptions).

We continue to present negative results on the consistent learnability of certain specific concept classes. For the rest of this section we only require that partial observations have *at most three* masked attributes. This suggests that the property of masking that compromises consistent learnability is not the frequency of the masked attributes, but rather the context in which they appear. Recall that Theorem 3.1 establishes that a similar property also compromises accurate predictability.

5.1 Intractability of Learning Parities

Our results so far leave open the possibility that every concept class PAC learnable by some hypothesis class is also consistently learnable from partial observations by the same hypothesis class. We dismiss this possibility by showing the concept class of parities, known to be properly PAC learnable [Helmbold *et al.*, 1992], not to be *properly* consistently learnable from partial observations, unless $\mathtt{RP} = \mathtt{NP}$.

Theorem 5.2 *The concept class $\mathcal{C}$ of parities over $\mathcal{A} \setminus \{x_t\}$ is not properly consistently learnable on x_t, unless $\mathtt{RP} = \mathtt{NP}$.*

The proof of Theorem 5.2 follows similar proofs from the literature (see, e.g., [Pitt and Valiant, 1988]). The reduction is from 3-SAT, and it 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. It is the case that with complete observations one can still force the learned hypothesis to depend on certain attributes, but the possible dependencies are necessarily restricted in a subtle, yet critical, manner.

5.2 Intractability of Learning Decision Lists

Rivest [1987] showed that the concept class of k-decision lists for a constant $k \in \mathbb{N}$ is properly PAC learnable, by showing how to identify a hypothesis that agrees with a given set of examples, and employing an Occam's Razor type argument [Blumer *et al.*, 1987]. He asked whether the same can be done when instead of examples one considers *partial* observations. In our notation, he defined *agreement*[3] of a formula φ with an observation $\mathtt{obs}$ to mean $\mathtt{val}(\varphi \,|\, \mathtt{obs}) = \mathtt{val}(x_t \,|\, \mathtt{obs})$, where x_t is the target attribute, which he assumed to be non-masked. As posed, the question almost always admits a trivial negative answer: an observation $\mathtt{obs}$ generally masks a set of examples such that the value of φ varies across examples, implying $\mathtt{val}(\varphi \,|\, \mathtt{obs}) = *$, and making φ disagree with $\mathtt{obs}$.

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 $\mathtt{obs}$ if φ does not conflict with the target attribute x_t under $\mathtt{obs}$. This weaker notion of "agreement" only requires that $\mathtt{val}(\varphi \,|\, \mathtt{obs}) = \mathtt{val}(x_t \,|\, \mathtt{obs})$ when $\mathtt{val}(\varphi \,|\, \mathtt{obs}), \mathtt{val}(x_t \,|\, \mathtt{obs}) \in \{0, 1\}$. 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 $\mathtt{RP} = \mathtt{NP}$. The negative answer carries to Rivest's original question, due to his stronger notion of "agreement".

Theorem 5.3 *The concept class $\mathcal{C}$ of monotone term 1-decision lists over $\mathcal{A} \setminus \{x_t\}$ is not properly consistently learnable on x_t, unless $\mathtt{RP} = \mathtt{NP}$.*

Theorem 5.3 establishes a separation from the framework of learning in the presence of random classification noise, in which the concept class of k-decision lists is known to be properly learnable for every constant $k \in \mathbb{N}$ [Kearns, 1998].

6 Related Work

Valiant [1984] recognizes early on the problem of missing information in observations, and proposes a model in which the target attribute is positive exactly when the target concept is positive in *all* examples masked by the observation. Some subsequent work follows a similarly-flavored approach in making certain assumptions regarding the value of the target attribute. Schuurmans and Greiner [1994] employ a masking process closely related to ours, although they assume that the target attribute is *never* masked. Their goal is also different, in that they focus on learning default concepts from partial observations, and not on the problem of recovering missing information. Goldman *et al.* [1997] consider a model in which the target attribute is positive/negative exactly when the target concept is correspondingly so in all the examples masked by the observation; the target attribute is masked only when the value of the target concept *cannot be deduced* from the non-masked attributes. Thus, they effectively treat "don't know" as a third distinguished value, and the problem reduces to that of learning ternary functions in what is essentially a complete information setting. In contrast, we focus on learning boolean formulas from partial observations without making such assumptions: the target attribute is masked arbitrarily, and when non-masked it simply indicates the value of the target concept for *some* example masked by the observation.

Decatur and Gennaro [1995] assume that attributes are masked independently, a *crucial* prerequisite for their statistical learning approach to work. We, on the other hand, consider the general setting where attributes are masked arbitrarily, staying thus closer to the spirit of the PAC semantics; arbitrary probability distributions model the unknown dependencies between properties of the environment, while arbitrary masking processes model the unknown dependencies on *what is observable* within the environment.

[3] Rivest [1987] used the term "consistency" rather than "agreement". We avoid, however, using the term "consistency" in this context, as it has a different meaning in our framework.

Multiple Instance Learning bears some resemblance to our work (see, e.g., [Dietterich *et al.*, 1997]). In that setting, a partial observation is an arbitrary bag of examples (usually assumed to be drawn independently), and an observation is positive exactly when at least one example is positive. Our partial observations can be seen as *structured* bags of (not independently drawn) examples. The structure restricts the possible combinations of examples, while the bag "labels" are much less informative, since they can assume either truth-value when bags contain both positive and negative examples.

Our masking process resembles the models of random attribute [Shackelford and Volper, 1988] and classification [Angluin and Laird, 1987] noise, where the values of attributes are affected, although independently across different examples, before being observed by an agent. Our model is perhaps closer to that of malicious noise [Valiant, 1985], in that no assumption is made as to how attributes are affected across examples. Malicious noise is known to render learnability almost impossible [Kearns and Li, 1993], while our more benign model, where affected attributes are explicitly made known to an agent, does not severely impair learnability.

7 Conclusions and Open Problems

We have presented an *autodidactic* learning framework as a means to recover missing information in data. Our framework builds on two natural generalizations of Valiant's PAC model [Valiant, 1984] to address the problem of learning from partial observations: learning consistently and predicting accurately. Producing accurate hypotheses was shown to be impossible under certain masking processes, even for computationally unbounded learners. On the positive side, producing consistent hypotheses was shown to be a *concrete* strategy for predicting as accurately as permitted by the masking process.

Within our framework we have presented a reduction technique, through which a number of natural boolean concept classes were shown to be consistently learnable. On the other hand, we have shown that properly consistently learning certain other concept classes is intractable, establishing, thus, a separation from the models of learning from complete or noisy observations. It remains open whether the intractability results are only representation-specific, and whether consistently learning from partial observations is (strictly) harder than learning under the random classification noise model.

A more general question concerns the consistent learnability of concept classes that are not shallow-monotone w.r.t. some efficiently computable set of substitutions. It is unclear whether such learnability results can be obtained through total reductions, which work for monotone concept classes.

Acknowledgments

The author is grateful to Leslie Valiant for his advice, and for valuable suggestions and remarks on this research.

References

[Angluin and Laird, 1987] Dana Angluin and Philip D. Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1987.

[Blumer *et al.*, 1987] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Occam's razor. *Information Processing Letters*, 24:377–380, 1987.

[Decatur and Gennaro, 1995] Scott E. Decatur and Rosario Gennaro. On learning from noisy and incomplete examples. In *Eighth Annual Conference on Computational Learning Theory*, pages 353–360, 1995.

[Dietterich *et al.*, 1997] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1–2):31–71, 1997.

[Goldman *et al.*, 1997] Sally A. Goldman, Stephen Kwek, and Stephen D. Scott. Learning from examples with unspecified attribute values (extended abstract). In *Tenth Annual Conference on Computational Learning Theory*, pages 231–242, 1997.

[Helmbold *et al.*, 1992] David Helmbold, Robert Sloan, and Manfred K. Warmuth. Learning integer lattices. *SIAM Journal on Computing*, 21(2):240–266, 1992.

[Kearns and Li, 1993] Michael J. Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, 1993.

[Kearns and Vazirani, 1994] Michael J. Kearns and Umesh V. Vazirani. *An Introduction to Computational Learning Theory*. The MIT Press, Cambridge, Massachusetts, U.S.A., 1994.

[Kearns, 1998] Michael J. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998.

[Pitt and Valiant, 1988] Leonard Pitt and Leslie G. Valiant. Computational limitations on learning from examples. *Journal of the ACM*, 35(4):965–984, 1988.

[Rivest, 1987] Ronald L. Rivest. Learning decision lists. *Machine Learning*, 2(3):229–246, 1987.

[Schafer and Graham, 2002] Joseph L. Schafer and John W. Graham. Missing data: Our view of the state of the art. *Journal of Psychological Methods*, 7(2):147–177, 2002.

[Schuurmans and Greiner, 1994] Dale Schuurmans and Russell Greiner. Learning default concepts. In *Tenth Canadian Conference on Artificial Intelligence*, pages 99–106, 1994.

[Shackelford and Volper, 1988] George Shackelford and Dennis Volper. Learning k-DNF with noise in the attributes. In *First Annual Workshop on Computational Learning Theory*, pages 97–103, 1988.

[Valiant, 1984] Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27:1134–1142, 1984.

[Valiant, 1985] Leslie G. Valiant. Learning disjunctions of conjunctions. In *Ninth International Joint Conference on Artificial Intelligence, Vol. 1*, pages 560–566, 1985.

[Valiant, 2000] Leslie G. Valiant. Robust logics. *Artificial Intelligence*, 117(2):231–253, 2000.