On Learning from Noisy and Incomplete Examples

Scott E. Decatur*
Aiken Computation Laboratory
Harvard University
Cambridge, MA 02138

Rosario Gennaro†
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

We investigate learnability in the PAC model when the data used for learning, attributes and labels, is either corrupted or incomplete. In order to prove our main results, we define a new complexity measure on statistical query (SQ) learning algorithms. The *view* of an SQ algorithm is the maximum over all queries in the algorithm, of the number of input bits on which the query depends. We show that a restricted view SQ algorithm for a class is a general sufficient condition for learnability in both the models of attribute noise and covered (or missing) attributes. We further show that since the algorithms in question are statistical, they can also simultaneously tolerate classification noise. Classes for which these results hold, and can therefore be learned with simultaneous attribute noise and classification noise, include k-DNF, k-term-DNF by DNF representations, conjunctions with few relevant variables, and over the uniform distribution, decision lists. These noise models are the first PAC models in which *all* training data, attributes and labels, may be corrupted by a random process.

Previous researchers had shown that the class of k-DNF is learnable with attribute noise if the attribute noise rate is known exactly. We show that all of our attribute noise learnability results, either with or without classification noise, also hold when the exact noise rate is not known, provided that the learner instead has a polynomially good *approximation* of the noise rate. In addition, we show that the results also hold when there is not just one noise rate, but a distinct noise rate for each attribute. Our results for learning with random covering do not require the learner to be told even an approximation of the covering rate and in addition hold in the setting with distinct covering rates for each attribute. Finally, we give lower bounds on the number of examples required for learning in the presence of attribute noise or covering.

1 Introduction

In this paper, we consider learning in the presence of faulty data in the "Probably Approximately Correct" (PAC) model introduced by Valiant [12]. In the PAC setting, a learner is given the task of determining a close approximation to an unknown $\{0,1\}$-valued target function. The learner is told a class of functions to which the target function belongs, and accuracy and confidence parameters ϵ and δ. The learner gains information about the target function by viewing labelled examples: a labelled example is composed of attributes and the example's label according to the target function. The learner is required to output an hypothesis such that, with high confidence, the accuracy of the hypothesis is high.

One limitation of the standard PAC model is that the data presented to the learner is assumed to be *noise-free*. In fact, most of the standard PAC learning algorithms would fail if even a small number of the labelled examples given to the learning algorithm were "noisy." While researchers have attempted to model different types of noise, in this work, we focus on noise affecting the attributes of the example. After discussing techniques for tolerating noise in the attributes of the example itself, we also address the simultaneous additional tolerance of noise in the label of the example.

The formal PAC model of attribute noise was simultaneously introduced by Shackelford and Volper [10] and Sloan [11]. In this model, each attribute of the example

*Author was supported by an NDSEG Fellowship and by NSF Grant CCR-92-00884. Author's email address: `sed@das.harvard.edu`

†Author was supported by NSF Grant 9121466-CCR and a graduate fellowship from the Consiglio Nazionale delle Ricerche, Italy. Author's email address: `rosario@theory.lcs.mit.edu`

Boolean vector is flipped randomly and independently with probability $\nu < 1/2$ before it is given to the learner. Shackelford and Volper showed that the class of k-DNF functions can be learned with any attribute noise rate $\nu < 1/2$ provided that the noise rate ν is known to the learner. Yet, Sloan showed that a reasonable algorithm, the one which chooses the function from the class with the fewest disagreements on a sample, cannot learn any *instance distinct* class (k-DNF is instance distinct) with an attribute noise rate greater than ϵ. Thus, the techniques used by Shackelford and Volper somehow avoid the pitfalls of simply "minimizing disagreements."

We generalize the result of Shackelford and Volper by showing a general sufficient condition for learnability in the presence of attribute noise. We make use of a recently developed tool for creating noise-tolerant algorithms: the statistical query (SQ) learning model [8]. In this model, instead of asking for labelled examples, the algorithm asks for estimates of the values of statistics defined over the distribution of labelled examples. This model may be viewed as a restriction on the way an algorithm uses the PAC example oracle since the example oracle could be used to simulate these statistical queries. Yet, this restriction has been found to be quite mild in that almost every class of concepts which is learnable in the PAC model is also learnable by statistical queries [8]. The most important use of the SQ model stems from the property that statistical queries can actually be simulated with the use of *noisy* example oracles. Specifically, an SQ algorithm can be simulated in the PAC model in the presence of classification noise, malicious errors, and even hybrid models combining these two types of noise [8, 5, 6].

In order to relate statistical query learning to PAC learning with attribute noise, we define a new complexity measure on statistical query algorithms called the *view*. The view of an SQ algorithm is the maximum over all queries in the algorithm, of the number of input variables on which the query depends. We show that any class that has an SQ algorithm with constant view can be learned in the presence of attribute noise arbitrarily approaching $\nu = 1/2$. We also show that any class that has an SQ algorithm with logarithmic view can be learned in the presence of attribute noise for any fixed constant noise rate $\nu < 1/2$.

We demonstrate that many commonly studied classes have constant or logarithmic view SQ algorithms. Examples include: (1) k-DNF (and k-CNF); (2) k-term DNF by DNF hypotheses (by converting the Blum and Singh [4] PAC algorithm); (3) conjunctions with few relevant variables; and (4) decision lists on the uniform distribution.

Although Goldman and Sloan [7] show that the very simple class of Boolean conjunctions can be learned in the presence of attribute noise without knowning the noise rate, the Shackelford and Volper result for k-DNF requires that the noise rate be specified to the learner. As it is much more plausible to expect a learner to have a reasonable *estimate* of a noise rate than to know it exactly, we show that all of our results hold when the learner is given, or can obtain, an estimate of the noise rate. We show that the accuracy of the estimate need only be within an inverse polynomial of the true noise rate.

Another negative result for attribute noise was shown by Goldman and Sloan [7] in which they consider a model of product attribute noise where each attribute x_i may have a distinct noise rate ν_i. They show that if only given an upper bound on all of the noise rates, it is impossible to learn any distinct (most common Boolean functions are also distinct) class if the upper bound is at least 2ϵ.

We show that all of our results hold in the product attribute noise setting, once again provided that the individual noise rates (or estimates of them) are given to or can be obtained by the learner. Thus, the Goldman and Sloan result for the hardness of product attribute noise relies not just on different noise rates, but specifically, a lack of knowledge of the noise rates.

Recall that our results hold for any class with a restricted view statistical query algorithm. We therefore may capitalize on the fact that the algorithms are statistical in nature, in order to show that they also intrinsicly tolerate classification noise, *i.e.* random noise on the *label* of the example. Specifically, we show that all of our attribute noise results above also hold in the presence of simultaneous attribute noise *and* classification noise. Such a hybrid model of learning with attribute noise and classification noise is the first model to capture the learning environment in which *all* of the training data, not just the attributes and not just the label, is corrupted by a random noise process.

In addition to the attribute noise model in which data is corrupted, we also investigate a model of attribute (and label) "covering" in which some of the information in a labelled example is missing (cf. Quinlan [9]). This model captures a learning environment in which a random process corrupts all aspects of the data, but errors can be detected although not corrected. Formally, we consider an example oracle which hides the value of each attribute (or label) independently with probability $\rho < 1$. We show that restricted view SQ algorithms are also sufficient for learning in this model. Specifically, we show that constant view SQ algorithms yield learnability with random covering for any rate ρ less than the information-theoretic limit of $\rho = 1$. Logarithmic view SQ algorithms yield learnability with random covering for any fixed constant rate $\rho < 1$.

A closely related model to restricted view SQ is the model of learning from *examples* with "restricted focus of attention" [2, 3] in which the learning algorithm submits functions to the example oracle which project the examples down to a smaller space. Although this model and restricted view SQ are not equivalent in general, they are equivalent within polynomial factors for the range of the view in which our results hold, and therefore our results hold in this model when the projectors are functions of constant or logarithmic numbers of attributes.

We note that our results can also be used to learn in a model in which data for learning may be both partially incorrect and partial incomplete, *e.g.* a model where each attribute or label may be either correct (with probability $1 - \rho - \nu$), incorrect (ν) or missing (ρ).

Finally, we show lower bounds on the number of examples required for learning in the presence of either attribute noise or attribute covering. Specifically, we show lower bounds for learning conjunctions whose dependence on the noise rates match the upper bounds achieved by our SQ simulations.

The remainder of the paper is organized as follows. In Section 2, we describe the relevant learning models. In Section 3, we present our main results demonstrating the relationship between learning with attribute noise and statistical query learning. We show in Section 4 how to simultaneously tolerate attribute noise and classification noise. In Section 5, we demonstrate the relationship between learning with random covering and statistical query learning. In Section 6, we apply the results of the previous sections by determining the "view" of the statistical query algorithms for various concept classes. In Section 7, we give lower bounds for learning conjunctions in the two noise models.

2 Learning Models

In this section we give formal definitions of the learning models and terminology used throughout the paper.

2.1 Noise-free and Noise-tolerant PAC Learning

An instance of the problem "PAC learning class $\mathcal{F}$" entails a hidden $\{0,1\}$-valued target function f chosen from the class $\mathcal{F}$, a hidden distribution D over X (the domain of f) and accuracy and confidence parameters ϵ and δ. The goal of a learning algorithm is to output, with probability at least $1 - \delta$, an hypothesis h with error at most ϵ. The error of an hypothesis is the probability[1] of drawing an *instance* (domain element) $x \in X$ on which $h(x) \neq f(x)$. An hypothesis with error at most ϵ is called ϵ-good. The learner gains information about f and D by interacting with an example oracle $EX(f, D)$ which on each requests returns in unit time an instance x (drawn according to D) and the classification $f(x)$ (we generally refer to an instance-label pair $\langle x, l \rangle$ as an *example* or *labelled example*). A learning algorithm is said to "PAC learn $\mathcal{F}$" if for any learning instance $\langle f, D, \epsilon, \delta \rangle$, with high probability, the algorithm outputs an ϵ-good hypothesis in time polynomial in $1/\epsilon$, $1/\delta$ and n (the length of instances in the domain of f).

In the *attribute noise* variant of PAC learning, the learning algorithm no longer has access to $EX(f, D)$, but instead has access to $EX^{\nu}_{\mathrm{ATT}}(f, D)$, where the parameter $\nu < 1/2$ is the noise rate. On each request, this new oracle draws an instance x according to D and computes its classification $f(x)$. Then each bit of the instance x is flipped independently with probability ν to create a new instance x' and the oracle returns $\langle x', f(x) \rangle$. We usually consider the case in which the learner is told the noise rate or an estimate of it. The learner is allowed to run in time polynomial in $1/(1 - 2\nu)$ and the standard parameters, but is still required to output an hypothesis which is ϵ-good with respect to *noise-free* data.

In the *random covering* variant of PAC learning, the learning algorithm no longer has access to $EX(f, D)$, but instead has access to $EX^{\rho}_{\mathrm{RC}}(f, D)$, where the parameter $\rho < 1$ is the cover rate. On each request, this new oracle draws an instance x according to D and computes its classification $f(x)$. Then each component of the instance x is covered (changed to "?") independently with probability ρ to create a new instance x' and the oracle returns $\langle x', f(x) \rangle$. The learner is allowed to run in time polynomial in $1/(1 - \rho)$ and the standard parameters, but is still required to output an hypothesis which is ϵ-good with respect to *noise-free* data. We assume for simplicity that the learner knows an upper bound on the cover rate ρ, however this assumption is not necessary since the true value of ρ can be very accurately estimated by the learner given access to EX^{ρ}_{RC}. Note that although attribute noise is only well defined over Boolean instance spaces, random covering may be considered for non-Boolean domains such as n-dimensional Euclidean space.

We will also consider the *classification noise* model. In this case the learning algorithm has access to the oracle to $EX^{\eta}_{\mathrm{CN}}(f, D)$, where the parameter $\eta < 1/2$ is the noise rate. On each request, this new oracle draws an instance x according to D and computes its label $l = f(x)$. It then returns $\langle x, l' \rangle$ where $l' = l$ with probability $1 - \eta$ and $l' = \neg l$ with probability η. We usually consider the case in which the learner is told the noise rate or an estimate of it. The learner is allowed to run in time polynomial in $1/(1 - 2\eta)$ and the standard parameters, but is still required to output an hypothesis which is ϵ-good with respect to *noise-free* data.

The final noise model we consider is the natural combination of classification noise and attribute noise in which examples are corrupted in either the attributes or the label. The learner has access to an oracle $EX^{\eta,\nu}_{\mathrm{CAT}}(f, D)$ that on each request draws an instance x according to the distribution D and computes its label $l = f(x)$. Then both the attributes (with probability ν) and the label (with probability η) are flipped independently and the learner is presented with the resulting pair $\langle x' l' \rangle$.

2.2 Statistical Query Learning

In the SQ model, the example oracle $EX(f, D)$ from the standard PAC model is replaced by a statistics oracle

[1] The probability is with respect to D and any randomization in the hypothesis h.

$STAT(f,D)$. An SQ algorithm queries the $STAT$ oracle for the values of various statistics on the distribution of labelled examples (*e.g.* "What is the probability that a randomly chosen labelled example $\langle x, l\rangle$ has variable $x_i = 0$ and $l = 1$?"), and the $STAT$ oracle returns the requested statistics to within some specified additive error. Formally, a statistical query is of the form $[\chi, \tau]$. Here χ is a mapping from labelled examples to $\{0,1\}$ (*i.e.* $\chi : X \times \{0,1\} \to \{0,1\}$) corresponding to an indicator function for those labelled examples about which statistics are to be gathered, while τ is an additive error parameter. A call $[\chi, \tau]$ to $STAT(f,D)$ returns an *estimate* $\hat{P}_\chi$ of $P_\chi = \Pr_D[\chi(x, f(x))]$ which satisfies $|\hat{P}_\chi - P_\chi| \le \tau$. The goal of the SQ algorithm is to output, with probability $1-\delta$, an ϵ-good hypothesis. An SQ learning algorithm is said to be polynomially efficient (in $1/\epsilon$, $1/\delta$, and n) if both (1) for every $[\chi,\tau]$, $1/\tau$ and the time required to evaluate χ is polynomially bounded, and (2) the running time of the SQ algorithm is polynomially bounded.

An SQ learning algorithm is said to use *query space* $\mathcal{Q}$ if it only makes queries of the form $[\chi,\tau]$ where $\chi \in \mathcal{Q}$. For every $\chi \in \mathcal{Q}$, we define the *view* of χ, denoted $|\chi|$, as the number of attributes (bits in the Boolean case) of the example x on which χ depends. We define the view of an SQ algorithm as the maximum over $\chi \in \mathcal{Q}$ of $|\chi|$. An algorithm is said to have *constant view* if its view is a constant independent of the input parameters n, ϵ and δ, while it is said to have *logarithmic view* if its view is bounded logarithmically in these parameters, *i.e.* $O(\log(\frac{n}{\epsilon\delta}))$.

3 Attribute Noise Learning

In this section we demonstrate the use of restricted view SQ algorithms for learning in the presence of attribute noise with a known noise rate ν.

Theorem 1 *If $\mathcal{F}$ is SQ learnable with view w, then $\mathcal{F}$ is PAC learnable with known attribute noise rate $\nu < 1/2$ in time polynomial in $(1-2\nu)^{-w}$ and 2^w.*

Note: The time bound is polynomial in the standard PAC parameters and $1/(1-2\nu)$ in either of two cases: (1) Any noise rate arbitrarily close to 1/2 for classes with constant view SQ algorithms, and (2) Any fixed constant noise rate less than 1/2 for classes with logarithmic (in the standard PAC parameters) view SQ algorithms.

Proof: For each query of the form $[\chi,\tau]$, we simulate the $STAT$ oracle by returning an estimate $\hat{P}_\chi$ of P_χ. Without loss of generality, consider a query χ with view w. For each $i \in I = \{0,\ldots,2^w-1\}$, let k_i be the n-bit vector with 0's in all components not viewed by χ, while the components viewed by χ are assigned bit values such that when these w bits are read in the canonical order on components, they represent i in binary. Define $\chi_i(x,b)$ for $i \in I$ to be $\chi(x \oplus k_i, b)$. Note that $\chi = \chi_0$. Recall that P_χ is the probability that a labelled example drawn from $EX(f,D)$ satisfies χ. Let P_χ^ν be the probability that a labelled example drawn from $EX^\nu_{\mathrm{ATT}}(f,D)$ satisfies χ. In order to relate the noise-free and noisy probabilities, we define matrix $A = \{a_{ij}\}$ where

$$a_{ij} = (1-\nu)^{w-H(i,j)}\nu^{H(i,j)} \qquad (1)$$

and $H(i,j)$ is the Hamming distance between k_i and k_j. If we define

$$q = \begin{bmatrix} P_{\chi_0} \\ P_{\chi_1} \\ \vdots \\ P_{\chi_{2^w-1}} \end{bmatrix} \quad \text{and} \quad q^\nu = \begin{bmatrix} P^\nu_{\chi_0} \\ P^\nu_{\chi_1} \\ \vdots \\ P^\nu_{\chi_{2^w-1}} \end{bmatrix}$$

then we have

$$Aq = q^\nu.$$

This last equation follows by noting that the probability that an example drawn from EX^ν_{ATT} satisfies χ is exactly the sum, over all possible corruptions of the w inputs bits of χ, of the probability of such a corruption times the probability of such a corrupt input satisfying χ.

Thus, in order to determine the noise-free probabilities in q from the noisy probabilities in q^ν, we simply note that

$$q = A^{-1}q^\nu.$$

It is not hard to verify (we in fact do so in the Appendix) that $A^{-1} = \{a^{(-1)}_{ij}\}$ specified below is indeed the inverse of A:

$$a^{(-1)}_{ij} = \frac{(\nu-1)^{w-H(i,j)}\nu^{H(i,j)}}{(2\nu-1)^w}. \qquad (2)$$

We may therefore draw a sample from EX^ν_{ATT} and use it to determine an estimate $\hat{q}^\nu$ of q^ν. We compute an estimate $\hat{q}$ of q as follows:

$$\hat{q} = A^{-1}\hat{q}^\nu.$$

By standard linear algebra, the error vector of $\hat{q}$, $e_q = \hat{q} - q$, then satisfies:

$$|e_q|_\infty \le |A^{-1}|_\infty \cdot |e_{q^\nu}|_\infty$$

where the infinity norms of column vectors and matrices are defined as follows:

$$|\vec{v}|_\infty = \sum_i |v_i|$$

$$|M|_\infty = \max_i \sum_j |a_{ij}|.$$

By drawing a sample of size polynomial in $1/\tau'$ (and w and $1/\delta$), we guarantee with high probability that for $e_{q^\nu} = \hat{q}^\nu - q^\nu$,

$$|e_{q^\nu}|_\infty \le \tau'.$$

We therefore need to determine $|A^{-1}|_\infty$. For all i,

$$\begin{aligned} \sum_j |a^{(-1)}_{ij}| &= \sum_j \left| \frac{(\nu-1)^{w-H(i,j)}\nu^{H(i,j)}}{(2\nu-1)^w} \right| \\ &= \frac{1}{(1-2\nu)^w} \sum_{h=0}^{w} \binom{w}{h} (1-\nu)^{w-h}\nu^h \\ &= \frac{1}{(1-2\nu)^w}. \end{aligned}$$

Thus

$$\begin{aligned} |e_q|_\infty &\leq |A^{-1}|_\infty \cdot |e_{q^\nu}|_\infty \\ &\leq \frac{1}{(1-2\nu)^w} \cdot \tau' \end{aligned}$$

Therefore, to achieve worst case additive difference τ in the vector $\hat{q}$, it is sufficient to sample the noisy oracle with tolerance $\tau' = \tau(1-2\nu)^w$. □

We note that the above theorem can be modified to allow learning in the presence of attribute noise in which for each attribute there is a distinct noise rate $\nu_i < 1/2$, provided that the learner is told the set of values $\{\nu_i\}$. This can be seen by inspecting the proof of Theorem 1 and noting that the crucial requirement is knowing the correct "noise producing matrix" which is possible with multiple noise rates, provided that the learner knows the noise rates.

Corollary 2 *If $\mathcal{F}$ is SQ learnable with view w, then $\mathcal{F}$ is PAC learnable with distinct known attribute noise rates $\{\nu_i < 1/2\}$ in time polynomial in $\prod_M (1-2\nu_i)^{-1}$ and 2^w, where M is the set of indices of the w largest noise rates.*

We next demonstrate that the noise rate need not be known exactly in order to learn. We quantify the slack allowed in the learner's knowledge of the noise rate in the following theorem whose proof is sketched in the Appendix. In addition we state the corollary which concerns learning with approximations of distinct noise rates.

Theorem 3 *If $\mathcal{F}$ is SQ learnable with view w and tolerance τ, then $\mathcal{F}$ is PAC learnable with attribute noise rate $\nu < 1/2$ in time polynomial in $(1-2\nu)^{-w}$ and 2^w, given an estimate $\hat{\nu}$ of the noise rate such that $|\nu - \hat{\nu}| = O(\frac{\tau(1/2-\nu)^{2w}}{w})$.*

Corollary 4 *If $\mathcal{F}$ is SQ learnable with view w and tolerance τ, then $\mathcal{F}$ is PAC learnable with distinct attribute noise rates $\{\nu_i < 1/2\}$ in time polynomial in $\prod_{i\in M}(1-2\nu_i)^{-1}$ and 2^w, given estimates $\{\hat{\nu}_i\}$ of the noise rates such that $\sum_{i\in N} |\nu_i - \hat{\nu}_i| = O(\tau(1/2-\nu)^{2w})$, where M is the set of indices of the w largest noise rates and N is the set of indices of the w worst approximations.*

It is interesting to contrast these results with the negative result of Goldman and Sloan [7] which states that if distinct noise rates are allowed and the learner is only given a single *upper bound* ν_b on all of the noise rates, then learning is impossible for any $\nu_b \geq 2\epsilon$. Therefore, the hardness in this case stems from the lack of knowledge of the amount of noise, not just the discrepancy between noise rates.

4 Tolerating Attribute Noise and Classification Noise

In this section we show that a restricted view SQ algorithm not only provides attribute noise tolerance, but simultaneous tolerance of both attribute noise and classification noise.

Theorem 5 *If $\mathcal{F}$ is SQ learnable with view w, then $\mathcal{F}$ is PAC learnable with known attribute noise rate $\nu < 1/2$ and known classification noise rate $\eta < 1/2$ in time polynomial in $(1-2\nu)^{-w}$, 2^w and $(1-2\eta)^{-1}$.*

Proof: Let $EX^{\eta,\nu}_{\mathrm{CAT}}(f, D)$ be an example oracle which when polled, draws an example x according to D, labels it according to f, and then flips each component of x with probability ν and flips the label with probability η. We combat such an example oracle by "undoing" both types of noise separately.

We combine techniques of Aslam and Decatur [1] with those used in the proof of Theorem 1. If we define $\overline{\chi}(x,b) = \chi(x, \neg b)$ and let $P^{\eta,\nu}_\chi$ be the probability that a labelled example drawn from $EX^{\eta,\nu}_{\mathrm{CAT}}(f, D)$ satisfies χ, then

$$P^{\eta,\nu}_\chi = (1-\eta)P^\nu_\chi + \eta P^\nu_{\overline{\chi}} \quad (3)$$

$$P^{\eta,\nu}_{\overline{\chi}} = (1-\eta)P^\nu_{\overline{\chi}} + \eta P^\nu_\chi \quad (4)$$

and therefore by solving the above equations for P^ν_χ:

$$P^\nu_\chi = \frac{(1-\eta)P^{\eta,\nu}_\chi - \eta P^{\eta,\nu}_{\overline{\chi}}}{1-2\eta}.$$

Thus, from (CAT) estimates $\hat{P}^{\eta,\nu}_\chi$ and $\hat{P}^{\eta,\nu}_{\overline{\chi}}$ of $P^{\eta,\nu}_\chi$ and $P^{\eta,\nu}_{\overline{\chi}}$, one can compute an (ATT) estimate $\hat{P}^\nu_\chi$ of P^ν_χ. The additive accuracy of the CAT estimates must be on the order of a $1-2\eta$ factor smaller than the desired additive accuracy of the ATT estimates. In the proof of Theorem 1, it was shown that in order to estimate noise-free queries with additive error τ, it is sufficient to estimate ATT queries (such as P^ν_χ) with additive error $\tau(1-2\nu)^w$. Therefore, the CAT estimates must be on the order of $\tau(1-2\nu)^w(1-2\eta)$. □

We can also prove a similar theorem regarding learnability in the presence of attribute noise and classification noise in which neither noise rate is known exactly, but both are known approximately.

Theorem 6 *If $\mathcal{F}$ is SQ learnable with view w, then $\mathcal{F}$ is PAC learnable with attribute noise rate $\nu < 1/2$ and classification noise rate $\eta < 1/2$ in time polynomial in $(1/2-\nu)^{-w}$, 2^w and $(1-2\eta)^{-1}$, given an estimate $\hat{\nu}$ of the attribute noise rate such that $|\nu - \hat{\nu}| = O(\frac{\tau(1/2-\nu)^{2w}}{w})$ and an estimate $\hat{\eta}$ of the classification noise rate such that $|\eta - \hat{\eta}| = O(\tau(1-2\nu)^w(1-2\eta))$.*

5 Attribute (and Label) Covering

Our goal in this section will be to establish a connection between restricted view SQ-learning and PAC learning with random covering. We will show that under certain conditions we can use SQ-algorithms to efficiently learn a concept when interacting with a faulty oracle EX^{ρ}_{RC}. The following lemma shows the connection between the cover rate ρ and the size of the sample needed to evaluate a statistical query with view w. This lemma follows by the independence of the covering and the fact that only w attributes need to be viewed. The lemma does not attempt to take advantage of queries in which knowing $< w$ attributes may uniquely determine the value of χ.

Lemma 7 *Given a query χ with view w the probability of drawing from EX^{ρ}_{RC} an example on which χ can be evaluated is at least $(1-\rho)^w$.*

Using the previous lemma we can relate SQ learnability to random covering learnability based on the view of the queries asked by the SQ algorithm. The proof of the following theorem is omitted in this abstract, and we note that the covering rate need not be specified to the learner, since it can be determined empirically.

Theorem 8 *If $\mathcal{F}$ is SQ learnable with queries of view w, then $\mathcal{F}$ is RC learnable in time polynomial in $(1-\rho)^{-w}$ and the standard PAC parameters.*

Note: It is trivial to extend these results to a model in which the label, in addition to the attributes, is covered randomly. Furthermore, although it may be more natural to consider this above extension, we note that our techniques also provide learnability in a model of learning with random covering of the attributes and *classification noise* affecting the label.

6 Classes with Restricted View SQ Algorithms

In this section, we state two theorems summarizing our results on PAC learnability in the presence of incomplete and noisy data. They apply to classes which are learnable with constant view and logarithmic view statistical query algorithms. We then enumerate the applications of these theorems by listing many classes with constant and logarithmic view statistical query algorithms.

Theorem 9 *If $\mathcal{F}$ is learnable by a constant view statistical query algorithm, then $\mathcal{F}$ is PAC learnable with:*

- *known (or approximate) attribute noise rate ν, for all $\nu < 1/2$.*
- *known (or approximate) classification noise rate η and known (or approximate) attribute noise rate ν, for all $\eta, \nu < 1/2$.*
- *(unknown) random covering rate ρ, for all $\rho < 1$.*
- *(unknown) classification noise rate η and (unknown) random covering rate ρ, for all $\eta < 1/2$ and $\rho < 1$.*

Theorem 10 *If $\mathcal{F}$ is learnable by a logarithmic view statistical query algorithm, then $\mathcal{F}$ is PAC learnable with:*

- *known (or approximate) attribute noise rate ν for all $\nu \leq \nu_0$, where ν_0 is any constant less than $1/2$.*
- *known (or approximate) classification noise rate η and known (or approximate) attribute noise rate ν for all $\eta < 1/2$ and $\nu \leq \nu_0$, where ν_0 is any constant less than $1/2$.*
- *(unknown) random covering rate ρ for all $\rho \leq \rho_0$, where ρ_0 is any constant less than 1.*
- *(unknown) classification noise rate η and (unknown) random covering rate ρ, for all $\eta < 1/2$ and $\rho \leq \rho_0$, where ρ_0 is any constant less than 1.*

We illuminate the applications of Theorems 9 and 10 with a listing of some concept classes which are learnable with constant and logarithmic view statistical query algorithms.

We first note that if we define a *w-term* to be a conjunction of at most w literals, then any subclass of DNF in which all functions are composed of terms from some polynomial sized set of w-terms is learnable by a w-view statistical query algorithm. The SQ algorithm simply performs "elimination" on the set of possible w-terms using the standard SQ algorithm for disjunctions in which the view of each query is the size of the term. By duality, analogous subclasses of CNF are also learnable. Thus, for constant w, there is a constant view SQ algorithm. Commonly studied classes with this property include k-DNF, k-CNF and constant depth k-decision trees. Since k-term-DNF is a subclass of k-CNF, it is also learnable by the k-CNF algorithm, and similarly for k-clause-CNF.

In order to learn k-term-DNF by a DNF representation with constant view, we make use of a result by Blum and Singh [4]. A careful conversion of their PAC algorithm yields an SQ algorithm with view k which learns k-term-DNF (k-clause-CNF) by DNF (CNF) representations.

One can also consider problems over non-Boolean instance domains in which we consider covering of attributes and labels and classification noise tolerance, but not attribute noise. One such problem with a constant view SQ algorithms is learning axis parallel rectangles over the domain $\Re^n$.

Corollary 11 *The problems of learning k-DNF, k-CNF, constant depth k-decision trees, k-term-DNF by DNF, k-clause-CNF by CNF, axis parallel rectangles over $\Re^n$ (no attribute noise applicability) can all be performed by constant view SQ algorithms and therefore in the PAC model in presence of noisy and incomplete data as specified by Theorem 9.*

Note: The class of conjunctions can actually be learned in the presence of attribute noise and classification noise without *any* knowledge of the attribute noise rate by combining our techniques with results of Goldman and Sloan [7].

When learning a function in which the number of relevant variables is much smaller than the total number of variables, it is desirable to have an algorithm with takes advantage of this property by using fewer example and/or less time. A common example is learning Boolean conjunctions over n variables, when the number of variables in the target is at most a constant k. The SQ algorithm for this problem [8] has view $k \log(1/\epsilon)$.

Under the uniform distribution on examples, decision lists can be approximated by $O(\log(1/\epsilon))$-length decision lists, and therefore can be learned by an SQ algorithm for this class. Such an algorithm has view $O(\log(1/\epsilon))$.

Corollary 12 *The problems of learning conjunctions of few relevant variables over any distribution and learning decision lists over the uniform distribution can performed by logarithmic view SQ algorithms and therefore in the PAC model in presence of noisy and incomplete data as specified by Theorem 10.*

7 Sample Complexity Lower Bounds

The simulations we give of SQ algorithms in the presence of attribute noise and random covering use more examples than the analogous simulations in a noise-free setting. The fact that there would be an increase in complexity due to the noise is not surprising. Here, we show that the algorithms for learning conjunctions, with either attribute noise or random covering, have optimal sample complexity dependence on the noise rates.

Proposition 1 *Learning conjunctions in the presence of attribute noise ν requires a sample complexity whose dependence on ν is $\Theta(\frac{1}{(1-2\nu)^2})$, while learning conjunctions in the presence of random covering ρ requires a sample complexity whose dependence on ρ is $\Theta(\frac{1}{1-\rho})$.*

Proof (Sketch): The upper bounds follow from the simulations of the standard SQ algorithm for conjunctions, while the lower bound for random covering follows directly from the number of examples required to see any given attribute even *once*.

For the attribute noise lower bound, we fix $n = 1$, $D(x = 0) = D(x = 1) = 1/2$, $f_1 = x$ and $f_2 = \overline{x}$. In the presence of noise, each example/label pair has probability within $(1 - 2\nu)/2$ of $1/4$, regardless of the target function. By lower bounding the tails of the binomial distribution, we can show that for $m = \frac{1}{c(1-2\nu)^2}$, there is a set of samples whose cumulative probability of being drawn is constant with respect to target f_1 such that every sample in this set has a probability with respect to f_2 at most a constant factor smaller than its probability with respect to f_1. If a sample from this set is drawn then the algorithm must guess between the two hypotheses. But since such a situation occurs with constant probability for either target, the guess (even if it hedges between the two) must have error at least 1/2 for at least one of the targets. $\square$

References

[1] Javed Aslam and Scott Decatur. Improved noise-tolerant learning and generalized statistical queries. Technical Report TR-17-94, Harvard University, July 1994.

[2] Shai Ben-David and Eli Dichterman. Learning with restricted focus of attention. In *Proceedings of the Sixth Annual ACM Workshop on Computational Learning Theory*, pages 287–296. ACM Press, 1993.

[3] Shai Ben-David and Eli Dichterman. Learnability with restricted focus of attention guarantees noise-tolerance. In *Proceedings of the Fifth International Workshop on Algorithmic Learning Theory*, 1994.

[4] Avrim Blum and Mona Singh. Learning functions of k terms. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 144–153, 1990.

[5] Scott Decatur. Statistical queries and faulty PAC oracles. In *Proceedings of the Sixth Annual ACM Workshop on Computational Learning Theory*, pages 262–268. ACM Press, July 1993.

[6] Scott Decatur. Learning in hybrid noise environments using statistical queries. In *Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*, pages 175–185, January 1995.

[7] Sally Goldman and Robert Sloan. Can PAC learning algorithms tolerate random attribute noise? Technical Report WUCS-92-25, Washington University, 1992.

[8] Michael Kearns. Efficient noise-tolerant learning from statistical queries. In *Proceedings of the 25^{th} Annual ACM Symposium on the Theory of Computing*, pages 392–401, San Diego, 1993.

[9] J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

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

[11] Robert Sloan. Types of noise in data for concept learning. In *Proceedings of the First Workshop on Computational Learning Theory*, pages 91–96. Morgan Kaufmann, 1988.

[12] Leslie Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, November 1984.

Appendix

Verifying the Inverse Noise Matrix: We show the correctness of the closed form specification of A^{-1} in Equation 2 by induction on w. In order to do so, we make the dependence on w explicit by denoting the error matrix for view w by A_w. For $w=1$, the correctness of A_1^{-1} can be verified by inspection:

$$A_1 = \begin{pmatrix} (1-\nu) & \nu \\ \nu & (1-\nu) \end{pmatrix} \quad \text{and}$$

$$A_1^{-1} = \begin{pmatrix} \frac{\nu-1}{2\nu-1} & \frac{\nu}{2\nu-1} \\ \frac{\nu}{2\nu-1} & \frac{\nu-1}{2\nu-1} \end{pmatrix}.$$

Note that the following inductive definitions of A_w and A_w^{-1} are equivalent to the closed form definitions:

$$A_w = \begin{pmatrix} (1-\nu)A_{w-1} & \nu A_{w-1} \\ \nu A_{w-1} & (1-\nu)A_{w-1} \end{pmatrix} \quad \text{and}$$

$$A_w^{-1} = \begin{pmatrix} \frac{\nu-1}{2\nu-1}A_{w-1}^{-1} & \frac{\nu}{2\nu-1}A_{w-1}^{-1} \\ \frac{\nu}{2\nu-1}A_{w-1}^{-1} & \frac{\nu-1}{2\nu-1}A_{w-1}^{-1} \end{pmatrix}.$$

The claim then follows by multiplying these matrices using the inductive hypothesis of $A_{w-1}^{-1}A_{w-1} = I$.

Proof Sketch of Theorem 3: Without loss of generality, we assume that the learner is given an an estimate $\hat{\nu} < 1/2$ of the noise rate ν such that $\hat{\nu} = (1+\alpha)\nu$. Let $\hat{A} = \{\hat{a}_{ij}\}$ be the error matrix parameterized by $\hat{\nu}$, *i.e.*

$$\hat{a}_{ij} = (1-\hat{\nu})^{w-k}\hat{\nu}^k \quad \text{where} \quad k = H(i,j).$$

The simulation solves the equation $\hat{A}\hat{q} = \hat{q}^\nu$, where the $\hat{q}^\nu$ is the empirical estimate of the noisy queries made with tolerance τ' (determined below) implying $|\hat{q}^\nu - q^\nu| < \tau'$. (All norms in this proof are ∞-norms.) Consider the following systems of equations:

- $Aq = q^\nu$**:** ideal system to solve using perfect estimates,
- $\hat{A}q_1 = q^\nu$**:** system using perfect estimate of q^ν, but approximate noise rate $\hat{\nu}$,
- $\hat{A}\hat{q} = \hat{q}^\nu$**:** actual system to solve, using estimates of q^ν and ν.

We are interested in determining the total error

$$e = |q - \hat{q}| \le |q - q_1| + |q_1 - \hat{q}|.$$

Solving the second two systems for q_1 and $\hat{q}$, we have $q_1 = \hat{A}^{-1}q^\nu$ and $\hat{q} = \hat{A}^{-1}\hat{q}^\nu$. Therefore the error due to the imperfect estimate of q^ν is

$$|q_1 - \hat{q}| \le |\hat{A}^{-1}||\hat{q}^\nu - q^\nu| < \frac{\tau'}{(1-2\hat{\nu})^w}.$$

Solving the first two systems for q and q_1, we have $q = A^{-1}q^\nu$ and $q_1 = \hat{A}^{-1}q^\nu$. Thus, the error due to the imperfect knowledge of ν is

$$|q - q_1| \le |q^\nu|(|A^{-1} - \hat{A}^{-1}|) \le (|A^{-1} - \hat{A}^{-1}|).$$

We therefore must bound $|A^{-1} - \hat{A}^{-1}|$. We write the matrix $\hat{A}^{-1} - A^{-1}$ as (c_{ij}) where:

$$c_{ij} = \hat{a}_{ij}^{(-1)} - a_{ij}^{(-1)} = \frac{(\hat{\nu}-1)^{w-k}\hat{\nu}^k}{(2\hat{\nu}-1)^w} - \frac{(\nu-1)^{w-k}\nu^k}{(2\nu-1)^w}.$$

By rearranging terms, and using the equivalence $\nu = \frac{\hat{\nu}}{(1+\alpha)}$, we have

$$c_{ij} = \hat{a}_{ij}^{(-1)}\left[1 - \frac{\left(1+\frac{\alpha}{1-\hat{\nu}}\right)^{w-k}}{\left(1+\frac{\alpha}{1-2\hat{\nu}}\right)^{w}}\right].$$

Since the multiplier of $\hat{a}_{ij}^{(-1)}$ is always positive, and $1+\frac{\alpha}{1-\hat{\nu}}$ and $1+\frac{\alpha}{1-2\hat{\nu}}$ are both greater than 1, we have

$$\begin{aligned} |c_{ij}| &\le |\hat{a}_{ij}^{(-1)}|\left[\left(1+\frac{\alpha}{1-2\hat{\nu}}\right)^w - 1\right] \\ &= \frac{|\hat{a}_{ij}^{(-1)}|}{(1-2\hat{\nu})^w}[(1-2\hat{\nu}+\alpha)^w - (1-2\hat{\nu})^w] \\ &\le \frac{|\hat{a}_{ij}^{(-1)}|\cdot\alpha w}{(1-2\hat{\nu})^w}. \end{aligned}$$

The final inequality holds since $1-2\hat{\nu}+\alpha$ and $1-2\hat{\nu}$ are both between 0 and 1. The norm of $|A^{-1} - \hat{A}^{-1}|$ is therefore

$$\begin{aligned} \sum_j |c_{ij}| &\le \frac{w\alpha}{(1-2\hat{\nu})^w}\cdot\sum_j |\hat{a}_{ij}^{(-1)}| \\ &= \frac{w\alpha}{(1-2\hat{\nu})^w}\cdot\frac{1}{(1-2\hat{\nu})^w} \\ &= \frac{w\alpha}{(1-2\hat{\nu})^{2w}}. \end{aligned}$$

The total error is then

$$|q-\hat{q}| \le \frac{\tau'}{(1-2\hat{\nu})^w} + \frac{w\alpha}{(1-2\hat{\nu})^{2w}}.$$

If the original SQ algorithm requires minimum tolerance τ, then the noisy estimates of queries can be made using minimum tolerance

$$\tau' \le \tau(1-2\hat{\nu})^w - \frac{w\alpha}{(1-2\hat{\nu})^w} = \Theta(\tau(1-2\hat{\nu})^w)$$

provided that the relative error of the noise rate estimate is at most $\alpha \le \frac{\tau(1-2\hat{\nu})^{2w}}{2w}$. For $\nu \ge 1/4$, this condition holds for $\alpha \le \frac{\tau(1-2\nu)^{2w}(1/2)^{2w}}{2w} = \frac{\tau(1/2-\nu)^{2w}}{2w}$. $\square$