Learning in Hybrid Noise Environments Using Statistical Queries

Scott Evan Decatur*
Harvard University

Abstract

We consider theoretical models of learning from noisy data. Specifically, we focus on learning in the *probability approximately correct* model as defined by Valiant. Two of the most widely studied models of noise in this setting have been *classification noise* and *malicious errors*. However, a more realistic model combining the two types of noise has not been formalized. We define a learning environment based on a natural combination of these two noise models. We first show that hypothesis testing is possible in this model. We next describe a simple technique for learning in this model, and then describe a more powerful technique based on statistical query learning. We show that the noise tolerance of this improved technique is roughly optimal with respect to the tolerance of the statistical query algorithm and that it provides a smooth tradeoff between the tolerable amounts of the two types of noise. Finally, we show that statistical query simulation yields learning algorithms for other combinations of noise models, thus demonstrating that statistical query specification truly captures the generic fault tolerance of a learning algorithm.

1 Introduction

An important goal of research in machine learning is to determine which tasks can be automated, and for those which can, to determine their information and computation requirements. One way to answer these questions is through the development and investigation of formal models of machine learning which capture the task of learning under plausible assumptions.

In this work, we consider the formal model of learning from examples called "probably approximately correct" (PAC) learning as defined by Valiant [14]. In this setting, a learner attempts to approximate an unknown target concept simply by viewing positive and negative examples of the concept. An adversary chooses, from some specified function class, a hidden $\{0,1\}$-valued target function defined over some specified domain of examples and chooses a distribution over this domain. The goal of the learner is to output in both polynomial time and with high probability, an hypothesis which is "close" to the target function with respect to the distribution of examples. The learner gains information about the target function and distribution by interacting with an example oracle. At each request by the learner, this oracle draws an example randomly according to the hidden distribution, labels it according to the hidden target function, and returns the labelled example to the learner. A class of functions $\mathcal{F}$ is said to be PAC learnable if there exists an algorithm which works for every target function in $\mathcal{F}$.

Whereas previous models required the learner to exactly determine the hidden concept but allowed the learner to use unbounded time, the PAC model requires the learner to work in efficient time, yet only requires the hypothesis returned by the learner to be "close" to the target concept.

*Aiken Computation Laboratory, Harvard University, Cambridge, MA 02138. E-mail: sed@das.harvard.edu.
Research supported by an NDSEG Doctoral Fellowship and by NSF grant CCR-92-00884.

These differences found in the PAC model seem to better reflect the requirements of learning in the real world. The PAC model has been widely adopted and there has been extensive research in providing algorithms and showing hardness results for this model [1].

However, one criticism of the PAC model is that the data used for learning is assumed to be noise free. In order to combat this deficiency, variations of PAC learning have been introduced which formalize the types of noise that might occur in a real training environment. Two of the most widely studied models of noise in computational learning theory have been *classification noise* [2] and *malicious errors* [15]. The classification noise model allows for common random occurrences (approaching 50% the time) of mislabelled examples, while the malicious error model allows for rare occurrences of adversarial corruption of the entire labelled example.

A useful tool for the development of algorithms in each of the above noise models has been another learning model, the statistical query (SQ) learning model [11]. A statistical query algorithm for a class can be very useful since it has been shown that it yields both a classification noise tolerant algorithm for the class [11] and a malicious error tolerant algorithm for the class [8]. In the SQ model, the learner may no longer view labelled examples, but instead may ask for estimates of the values of various statistics based on the distribution of labelled examples. Since such statistics could be accurately estimated with high probability by a large sample of labelled examples, one can view this model as restricting the way in which the learner may use the example oracle. Yet this restriction has been found to be quite mild, in that almost all classes which have PAC algorithms also have SQ algorithms [11]. Furthermore, these SQ algorithms can be easily derived from the corresponding PAC algorithms. Therefore, for all of these classes, noise tolerant algorithms exist in each of the two noise models described above.

Although many algorithms have been constructed to learn in the presence of each type of noise separately, a more realistic model combining classification noise and malicious errors has not been formalized. In this hybrid model, occasionally entire labelled examples would be corrupted, while the the remaining examples would have some fixed probability of being randomly misclassified. A noise model which combines these two types of data corruption is at least as hard to learn in as each separate noise model. In this paper, we define a learning environment based on a natural combination of these two noise models called *classification and malicious* error (CAM).

We first show that one can perform hypothesis testing in the presence of CAM error, *i.e.* use only noisy data to detect which hypothesis from a set of hypotheses has the smallest error on *noise-free* data. We then show how one can use existing techniques to construct algorithms for learning in this noise model. Specifically, we show how to take an algorithm which tolerates classification noise and add to it the ability to also tolerate malicious errors. We then derive the limits of this technique in terms of the amount of noise tolerance and note that optimal noise tolerance cannot be achieved by these methods.

We next show a different technique for generating CAM-tolerant learning algorithms which tolerate strictly more noise. This technique is based on simulating statistical query algorithms. We show that for *any* SQ algorithm, the classification noise tolerance achieved is optimal while the malicious error tolerance is within a logarithmic factor of optimal. In fact, many SQ algorithms yield optimal malicious error tolerance. Furthermore, this technique provides a smooth tradeoff between the amount of tolerable classification noise and malicious error.

Finally, we describe how the definition of a hybrid noise environment can be extended to combine various other types of noise. Learnability in these new hybrid noise models may also be achieved through the simulation of statistical query algorithms. This generality of the usefulness of statistical query specification demonstrates its ability to capture the fault tolerance intrinsic to a learning problem.

2 Learning Models

In Valiant's PAC model of learning from labelled examples [14], an adversary selects both the hidden target $\{0,1\}$-valued function f from a known class of functions $\mathcal{F}$ and the hidden distribution D which is defined over the domain of f. The domain of f constitutes the set of possible examples. This set is often the Boolean hypercube $\{0,1\}^n$, in which case n is the common length of all examples. The learner is given access to an example oracle $EX(f,D)$ which, when polled by the learner, returns $\langle x, f(x)\rangle$, an example x drawn randomly according to D and its correct labelling with respect to f. The learner is also given accuracy parameter $\varepsilon \in (0,1)$ and confidence parameter $\delta \in (0,1)$. In time polynomial in n, $1/\varepsilon$, and $1/\delta$, the learner must output an hypothesis h which, with probability at least $1-\delta$, has the following property: given an example x drawn randomly according to D, the probability that $f(x) \neq h(x)$ is at most ε. If h has this property, we say that it is ε-close to f on D.

We next describe two variants of PAC learning which model noise in the learning process. These learning models, in addition to the statistical query model described below, all differ from the standard PAC model in which oracle they use to interact with f and D. Yet, all models still require the learner to output, with probability at least $1-\delta$, an hypothesis h which is ε-close to f on D, *i.e.* with respect to noise-free labelled examples.

Angluin and Laird [2] introduced the model of PAC learning with *classification noise*, in which the learner has access to a noisy example oracle $EX^{\eta}_{\mathrm{CN}}(f,D)$. When a labelled example is requested from this oracle, an example is chosen according to the hidden distribution D, and returned. With probability $1-\eta$, the correct labelling of the example according f is returned, while with probability η, the incorrect classification is returned. The learner is given η_b, an upper bound on the noise rate, such that $0 \leq \eta \leq \eta_b < 1/2$. The running time of a learning algorithm is allowed to be polynomial in $\frac{1}{1/2-\eta_b}$ in addition to the usual parameters. We say that a class is learnable with classification noise if it is learnable for every classification noise rate η less than the information theoretic limit of $1/2$.

The PAC model with *malicious errors* was introduced by Valiant [15] and studied further by Kearns and Li [12]. In this model, the learner has access to an example oracle $EX^{\beta}_{\mathrm{MAL}}(f,D)$. When a labelled example is requested from this oracle, with probability $1-\beta$, an example is chosen according to the hidden distribution D, correctly labelled according to the hidden target concept f, and returned to the learner. However, with probability β, a malicious adversary selects any example it chooses and labels it either positive or negative. Kearns and Li [12] showed that for "distinct" classes (virtually all interesting classes are distinct), it is impossible to tolerate a malicious error rate of $\beta \geq \frac{\varepsilon}{1+\varepsilon}$.

The statistical query (SQ) model [11], a further variant of the noise-free PAC model, has been a useful tool in the construction of *noise-tolerant* PAC algorithms. While learning in the PAC model may be based on specific properties of individual examples, learning in the SQ model is based on statistical properties of large sets of examples. In the SQ model, the PAC example oracle $EX(f,D)$ is replaced by a statistics oracle $STAT(f,D)$. The learner interacts with $STAT(f,D)$ by asking it queries of the form $[\chi,\tau]$ where χ is a $\{0,1\}$-valued function on *labelled* examples and $\tau \in (0,1)$ is the *tolerance* of the query. The query is a request for the value P_χ, the probability that $\chi(x,l)=1$ when $\langle x,l\rangle$ is a labelled example drawn randomly from $EX(f,D)$. Thus, P_χ is the probability of drawing a labelled example that has "property" χ. The statistics oracle returns an approximation $\hat{P}_\chi$ such that $|\hat{P}_\chi - P_\chi| \leq \tau$.

The *query space* $\mathcal{Q}$ of an SQ algorithm is the set of all possible queries χ which the algorithm asks the statistics oracle on all possible runs. The *tolerance* of an SQ algorithm is the lower bound

on the tolerances of the queries the algorithm makes to the oracle. A class is said to be SQ learnable if: (1) there exists an SQ algorithm which makes only a polynomial number of queries, (2) there exists a polynomial bound on the time required to evaluate every χ used by the algorithm, and (3) there exists a polynomial bound on the inverse of the tolerance of every query made by the algorithm.

Kearns [11] has shown that if a class has an SQ algorithm, then it is learnable with any amount of classification noise $\eta < 1/2$. Decatur [8] has shown that if a class has an SQ algorithm with tolerance τ, then it is learnable with malicious error $\beta = \Theta(\tau)$.[1] Both results are based on simulations of SQ algorithms in the respective noise models. Despite the noise in the examples, the simulations are able to effectively reproduce a noise-free statistics oracle. The sample complexities of the noise-tolerant PAC algorithms depend on the number of queries, the tolerance and the query space of the SQ algorithm.

We next examine a hybrid model of noise based on classification noise and malicious errors.

3 Learning with Classification Noise *and* Malicious Errors

We first define the new learning model which combines the two noise types and describe how to perform hypothesis testing in it. We then give a simple construction to learn in this model. Next, we give an improved technique for CAM learning based on simulating statistical queries. We prove that the noise tolerance of this technique is roughly optimal and observe that it provides a smooth tradeoff in tolerance between the two types of errors.

3.1 The Hybrid Noise Model

We formally define *classification and malicious* error (CAM) as a PAC variant using an example oracle with the following behavior:

$$EX^{\eta,\beta}_{\mathrm{CAM}}(f,D) = \begin{cases} EX(f,D) & 1-\eta-\beta \\ EX(\neg f,D) & \eta \\ \text{Adversary} & \beta \end{cases}$$

The learner is told η_b such that $\eta \leq \eta_b < \frac{1}{2}$ and given $EX^{\eta,\beta}_{\mathrm{CAM}}(f,D)$. In time polynomial in $1/\varepsilon$, $1/\delta$, n, and $\frac{1}{1/2-\eta_b}$, the learner must output an hypothesis h which with probability at least $1-\delta$ is ε-close to f on D.

3.2 Hypothesis Testing using a CAM Example Oracle

In this section, we show how to use the $EX^{\eta,\beta}_{\mathrm{CAM}}$ oracle to perform so called "hypothesis testing" in order to determine which hypotheses of a given set have relatively small error with respect to the *noise-free* example oracle. We first prove a theorem which states an upper bound on the number of labelled examples from a CAM oracle sufficient to rank two hypotheses with different error rates.

[1] We often use $O, \Omega, \Theta, o, \omega$ when characterizing the asymptotic behavior of functions of variables which approach 0 as opposed to the standard usage in which variables approach ∞. For instance, since τ approaches 0, we use $O(\tau)$ to denote a function g for which there exists constants k and τ_0 such that for all $\tau \leq \tau_0$, $g(\tau) \leq k\tau$. We also make use of "soft" order notation when we are not concerned with lower order logarithmic factors. Specifically, when $b > 1$, we define $\tilde{O}(b)$ to mean $O(b\log^c b)$ for some constant $c \geq 0$. When $b < 1$, we define $\tilde{O}(b)$ to mean $O(b\log^c(1/b))$ for some constant $c \geq 0$. We define $\tilde{\Omega}$ similarly for some constant $c \leq 0$ and analogously define $\tilde{\Theta}, \tilde{o}, \tilde{\omega}$. When discussing asymptotics related to η, we consider $\eta \to 1/2$, or correspondingly $(1/2-\eta)^{-1} \to \infty$.

Theorem 1 *Let h_1 and h_2 be two hypotheses with error rates ε_1 and ε_2 such that $\varepsilon_2 - \varepsilon_1 = \gamma > 0$. If $\beta \leq \gamma(1/2-\eta)/2$, then a sample S of size $O\left(\frac{(\varepsilon_1+\gamma)\log(1/\delta)}{\gamma^2(1/2-\eta)^2}\right)$ drawn from $EX^{\eta,\beta}_{\mathrm{CAM}}(f,D)$ is sufficient to guarantee, with probability at least $1-\delta$, that h_1 disagrees with fewer labelled examples than h_2 in the sample S.*

Proof: By generalizing a technique of Laird [13], Aslam and Decatur [5] show that two important parameters determine the number of labelled examples sufficient to perform hypothesis testing by any type of example oracle. These parameters are (1) t — the probability of drawing a labelled example on which the two hypotheses disagree; and (2) η' — the conditional probability, given the hypotheses disagree with each other, that the hypothesis with smaller error disagrees with the label. They show that it is sufficient to draw a sample S of $O(\log(1/\delta) \cdot t^{-1}(1-2\eta')^{-2})$ labelled examples to ensure with probability at least $1-\delta$ that h_2 disagrees with more labelled examples in S than h_1 does. We simply derive the values of t and η' for $EX^{\eta,\beta}_{\mathrm{CAM}}$.

Let d be the probability (with respect to D and any randomization in h_1 or h_2) of drawing an example x according to D in which $h_1(x) = h_2(x) \neq f(x)$. Let t_1 be the probability (with respect to D and any randomization in h_1 or h_2) of drawing a labelled example $\langle x, l\rangle$ from $EX^{\eta,\beta}_{\mathrm{CAM}}(f,D)$ in which $h_1(x) \neq h_2(x) = l$.[2] Similarly, let t_2 be the probability (with respect to D and any randomization in h_1 or h_2) of drawing a labelled example $\langle x, l\rangle$ from $EX^{\eta,\beta}_{\mathrm{CAM}}(f,D)$ in which $h_2(x) \neq h_1(x) = l$. Then we have

$$\begin{aligned} t_1 &= \beta + (\varepsilon_1 - d)(1-\eta-\beta) + (\varepsilon_2 - d)\eta \\ t_2 &= (\varepsilon_2 - d)(1-\eta-\beta) + (\varepsilon_1 - d)\eta \end{aligned}$$

and since $t = t_1 + t_2$, we have

$$t = \beta + (\varepsilon_1 + \varepsilon_2 - 2d)(1-\beta).$$

The value of η' is simply t_1/t and therefore

$$\eta' = \frac{\beta + (\varepsilon_1 - d)(1-\eta-\beta) + (\varepsilon_2 - d)\eta}{\beta + (\varepsilon_1 + \varepsilon_2 - 2d)(1-\beta)}.$$

From the above values for t and η', and by our assumption that $\beta \leq \gamma(1-2\eta)/4$, we have

$$\begin{aligned} t^{-1}(1-2\eta')^{-2} &= \frac{\beta + (\varepsilon_1 + \varepsilon_2 - 2d)(1-\beta)}{[(\varepsilon_2 - \varepsilon_1)(1-\beta-2\eta) - \beta]^2} \\ &\leq \frac{\gamma(1-2\eta)/4 + (\varepsilon_1 + \varepsilon_2)}{[\gamma(1-2\eta)/2]^2} \\ &= \frac{\gamma(1-2\eta)/4 + (\gamma + 2\varepsilon_1)}{\gamma^2(1/2-\eta)^2}. \\ &= O\left(\frac{\varepsilon_1 + \gamma}{\gamma^2(1/2-\eta)^2}\right). \end{aligned}$$

□

A standard application of a result such as Theorem 1 is that it allows one to select, with high probability, an hypothesis with error at most ε from a set of hypotheses containing at least one with error at most $\varepsilon/2$. Specifically, consider a set of N hypothesis, in which one of these hypotheses,

[2] Since this probability actually depends on a dynamically playing adversary, we consider the worst possible case in which the adversary, at every opportunity, returns a labelled example $\langle x, l\rangle$ such that $l = h_2(x) \neq h_1(x)$.

say h_1, is guaranteed to have error rate at most $\varepsilon/2$. By selecting the hypothesis from this set which has the smallest empirical error on a sufficiently large sample, we can be confident that the error rate of this hypothesis is no more than ε. For this procedure to work, it is enough for h_1 to have smaller empirical error than any hypothesis with true error more than ε. Thus, we allocate $\delta/(N-1)$ probability of failure to each comparison of h_1 with the other hypotheses and note that for each hypothesis of error more than ε, the corresponding gap is $\gamma \geq \varepsilon/2$.

Corollary 2 *Let $h_1, \ldots, h_N$ be hypotheses with error rates $\varepsilon_1 \leq \cdots \leq \varepsilon_N$ such that $\varepsilon_1 \leq \varepsilon/2$. If $\beta \leq \varepsilon(1/2-\eta)/4$, then a sample S of size $O\left(\frac{\log(N/\delta)}{\varepsilon(1/2-\eta)^2}\right)$ drawn from $EX^{\eta,\beta}_{\mathrm{CAM}}(f,D)$ is sufficient to guarantee, with probability at least $1-\delta$, that the hypothesis with the fewest disagreements on S has error no more than ε.*

3.3 Simple Strategies for CAM Learning

We next examine how existing tools for both classification noise and malicious error learning may be combined to achieve CAM learning. Specifically, we consider the strategy of starting with an algorithm which tolerates one type of noise, and transforming it so that it additionally tolerates the second type of noise.

Given an algorithm which tolerates malicious errors, we know of no transformation which adds the ability to tolerate classification noise. This is illustrated by the class of parity functions. The class of parity functions has a malicious error tolerant algorithm (yielded by the technique of "multiple-runs" [12] discussed below, on the noise-free algorithm for learning parity functions [10]), yet it is not known how to learn parity functions even in the presence of classification noise alone.

Conversely, such a transformation does exist when starting with an algorithm which tolerates classification noise. Kearns and Li [12] describe a technique which takes a PAC algorithm with sample complexity m which does not tolerate malicious errors and by running it many times, converts it into one which tolerates a malicious error rate $\beta = \Theta(\frac{\log m}{m})$. We can take an algorithm which tolerates classification noise, and use this technique to create one which tolerates CAM error. The amount of CAM error which can be tolerated using these techniques may be upper bounded by using lower bounds on the sample complexity of classification noise learning. Aslam and Decatur [5] have shown that any algorithm for function class $\mathcal{F}$ which tolerates classification noise must have sample complexity at least $m = \Omega(\frac{\mathrm{VCDim}(\mathcal{F})}{\varepsilon(1/2-\eta)^2})$.[3] In fact, most known classification noise algorithms have sample complexity at least $m = \Omega(\frac{\mathrm{VCDim}(\mathcal{F})}{\varepsilon^2(1/2-\eta)^2})$. Thus, these techniques yield algorithms which tolerate CAM with any $\eta < 1/2$, but β at most $\tilde{O}(\varepsilon(1/2-\eta)^2/\mathrm{VCDim}(\mathcal{F}))$. Specifically, the dependence of β on ε and η is at best $\tilde{O}(\varepsilon(1/2-\eta)^2)$ and often $\tilde{O}(\varepsilon^2(1/2-\eta)^2)$. We next show a different technique for deriving algorithms which tolerate more CAM error.

3.4 Simulating Statistical Queries for Improved CAM Learning

We show how to efficiently simulate an SQ algorithm in the PAC model in the presence of CAM error. The strategy is to perform a simulation similar to that used for classification noise alone, but to do so with sharper tolerances in order to also tolerate the malicious errors.

Theorem 3 *Given an SQ learning algorithm for $\mathcal{F}$ with minimum tolerance τ, one can construct a PAC learning algorithm for $\mathcal{F}$ which can tolerate CAM for all $\eta < 1/2$ and $\beta \leq \beta_* = \Theta(\tau(1/2-\eta))$.*

[3] The Vapnik-Chervonenkis dimension, or VCDim, of a class $\mathcal{F}$ is a combinatorial measure which is commonly used to characterize the required sample complexity for learning $\mathcal{F}$. Note that $\mathrm{VCDim}(\mathcal{F}) \geq 1$.

Note: We can tolerate as much classification noise in the CAM model as was possible in the standard classification noise model. In the case of malicious error, the amount of error tolerable depends on the tolerance of the SQ algorithm (as was the case in the model of malicious error alone) and the classification noise rate. This theorem provides a smooth tradeoff of classification noise against malicious errors in that as η approaches 0, the amount of malicious error tolerable approaches the level tolerable in the model of malicious error alone.

Note: Aslam and Decatur [4] show the dependence of τ on ε need never be less than $\varepsilon/\log(1/\varepsilon)$. Therefore the dependence of β on ε and η need *never* be worse than $\Omega(\frac{\varepsilon}{\log(1/\varepsilon)}(1/2-\eta))$ which improves upon the noise tolerance achieved in Section 3.3. In fact, the tolerance for many classes is $\tau = \Theta(\varepsilon/\text{VCDim}(\mathcal{F}))$ in which case the malicious error tolerable is $\beta = \Theta(\varepsilon(1/2-\eta)/\text{VCDim}(\mathcal{F}))$.

Proof: The simulation of the SQ algorithm draws a sample of labelled examples from the CAM example oracle, uses that sample to estimate all of the queries in the SQ algorithm, and outputs the hypothesis returned by the SQ algorithm. If the simulation of the *STAT* oracle is correct (*i.e.* within proper tolerances) for all queries made, then by the correctness of the SQ algorithm, the output hypothesis is ε-good. Therefore, with probability $1-\delta$, the simulation algorithm must answer all queries accurately.

The strategy to simulate queries is two-fold: bound the fraction of examples on which the malicious adversary plays and ensure that the remaining examples *very accurately* estimate quantities used to estimate P_χ in the presence of classification noise. The additional accuracy is required to accommodate error introduced by the malicious error examples. The quantities used to estimate P_χ are based on the *classification noise alone* simulation of SQ algorithms of Aslam and Decatur [6].

For each query $[\chi, \tau]$, we wish to estimate P_χ, the probability that a labelled example drawn from $EX(f,D)$ satisfies χ, to within $\pm\tau$. Let P_χ^τ be the probability that a labelled example drawn from $EX^{\eta,\beta}_{\text{CAM}}(f,D)$ satisfies χ given that the labelled example was not chosen by the malicious adversary. Let $P_{\bar{\chi}}$ be the probability that a labelled example drawn from $EX(f,D)$ satisfies χ if we negate the example's label. Define $P_{\bar{\chi}}^\tau$ similarly. These definitions yield the following equations:

$$P_\chi^\tau = \frac{(1-\beta-\eta)P_\chi + \eta P_{\bar{\chi}}}{1-\beta} \tag{1}$$

$$P_{\bar{\chi}}^\tau = \frac{(1-\beta-\eta)P_{\bar{\chi}} + \eta P_\chi}{1-\beta} \tag{2}$$

Solving Equation 2 for $P_{\bar{\chi}}$ and substituting into Equation 1, we get the following expression for P_χ:

$$P_\chi = \frac{(1-\beta-\eta)P_\chi^\tau - \eta P_{\bar{\chi}}^\tau}{1-\beta-2\eta} \tag{3}$$

Thus, in order to estimate P_χ, it is enough to have estimates of P_χ^τ, $P_{\bar{\chi}}^\tau$, η and β. We make use of the following lemma which states how accurately these quantities must be estimated or "guessed." The lemma is proven by the use of some simple algebra.

Lemma 4 *Provided $\beta \le \beta_* = \Theta(\tau(1/2-\eta))$, in order to estimate P_χ to within $\pm\tau$, it is sufficient to estimate P_χ^τ and $P_{\bar{\chi}}^\tau$ each to within $\pm\tau(1/2-\eta)/8$ and to guess an $\hat{\eta}$ such that $|\eta-\hat{\eta}| \le \tau(1/2-\eta)/8$.*

Proof: Omitted. □

We simulate the SQ algorithm multiple times, each with a different "guess" for η. Each simulation attempts to achieve error no more than $\varepsilon/2$. On any of the runs in which the noise rate

is guessed accurately enough, the hypothesis output is $\varepsilon/2$-good with high probability. By using a sliding scale for the noise rate guesses, it is possible to construct a set of $O(\frac{1}{\tau}\log\frac{1}{1/2-\eta_b})$ guesses such that for at least one of the guesses $\hat{\eta}$, we have $|\eta-\hat{\eta}| \leq \tau(1/2-\eta)/8$. Details of this construction are given in the full paper.

We compare the hypotheses generated by all of the different runs and output the one with smallest empirical error on a random sample. Corollary 2 can be used to show that this hypothesis is ε-good with high probability. Note that we allocate $\delta/2$ probability of failure to the hypothesis testing and $\delta/2$ probability of failure to the determination of the estimates within each run.

We next determine the sample size sufficient to ensure, with probability $1-\delta/2$, that every P^r_χ and $P^r_{\bar{\chi}}$ is estimated accurately. If we define $\tau' = \tau(1/2-\eta)/8$ and $\theta = \tau' - \beta$, then assuming $\beta \leq \beta_* = \tau(1/2-\eta)/16$, we have $\theta = \Omega(\tau(1/2-\eta))$. Let S be a sample of size m (to be determined below) labelled examples drawn from $EX^{\eta,\beta}_{\mathrm{CAM}}$, let S_1 be the subset of S chosen by the malicious adversary and let $S_2 = S \setminus S_1$, *i.e.* examples chosen according to the correct distribution but possibly mislabelled. Let S^χ be the subset of S which satisfies χ. Define S_1^χ and S_2^χ similarly. Let our estimate for P^r_χ be $|S^\chi|/|S|$. We show that if both Conditions 1 and 2 (given below) hold, then for all $\chi \in \mathcal{Q}$, $\left|P^r_\chi - |S^\chi|/|S|\right| \leq \tau'$. That is, the fraction of examples in the sample from $EX^{\eta,\beta}_{\mathrm{CAM}}$ which satisfy χ is a good estimate for P^r_χ.

Condition 1: $\dfrac{|S_1|}{|S|} \leq \beta + \dfrac{\theta}{2}$

Condition 2: $\forall \chi \in \mathcal{Q},\ \left|\dfrac{|S_2^\chi|}{|S_2|} - P^r_\chi\right| \leq \dfrac{\theta}{2}$

Condition 1 states that the malicious adversary played on no more than a $\beta + \frac{\theta}{2}$ fraction of the m examples. Condition 2 states that for all queries, the fraction of non-adversarially labelled examples satisfying χ is within $\frac{\theta}{2}$ of the expected fraction P^r_χ. In order to show these two conditions jointly imply accurate estimates, we simply verify that they imply $\forall \chi \in \mathcal{Q}$, $|S_\chi|/|S| \leq P^r_\chi + \tau'$ and $|S_\chi|/|S| \geq P^r_\chi - \tau'$:

$$\begin{aligned}
\frac{|S^\chi|}{|S|} = \frac{|S_1^\chi|}{|S|} + \frac{|S_2^\chi|}{|S|} &\leq \frac{|S_1|}{|S|} + \frac{|S_2^\chi|}{|S_2|} \\
&\leq \left(\beta + \frac{\theta}{2}\right) + \left(P^r_\chi + \frac{\theta}{2}\right) \\
&= P^r_\chi + \tau'.
\end{aligned}$$

$$\begin{aligned}
\frac{|S^\chi|}{|S|} \geq \frac{|S_2|}{|S|}\cdot\frac{|S_2^\chi|}{|S_2|} &\geq \left(1-\beta-\frac{\theta}{2}\right)\left(P^r_\chi - \frac{\theta}{2}\right) \\
&= P^r_\chi - \beta P^r_\chi - \frac{\theta}{2}\left[P^r_\chi + 1 - \beta - \frac{\theta}{2}\right] \\
&\geq P^r_\chi - \beta - \frac{\theta}{2}\cdot 2 \\
&= P^r_\chi - \tau'.
\end{aligned}$$

Condition 1 requires the empirical fraction of a set of independent trials of a Bernoulli random variable to be no more than $\theta/2$ larger than its expected value, while Condition 2 requires similar empirical fractions to be within $\pm\theta/2$ of their expected values. Using standard Chernoff bounds [3] and uniform convergence results [7], to ensure that (with probability $1-\delta/2$) both conditions hold, it is sufficient to use a sample of size $m = O(\frac{1}{\theta^2}\log\frac{|\mathcal{Q}|}{\delta})$ if the query space $\mathcal{Q}$ is finite or a sample

of size $m = O(\frac{q}{\theta^2}\log\frac{1}{\theta} + \frac{1}{\theta^2}\log\frac{1}{\delta})$ if $\mathcal{Q}$ is infinite with finite VC Dimension q. Alternatively, one could use a separate sample for each of the N queries made by the algorithm, thus using a total of $m = O(\frac{N}{\theta^2}\log\frac{N}{\delta})$ examples. □

3.5 Optimality of CAM Tolerance

Theorem 3 states that *any* class which is SQ learnable is PAC learnable with CAM for all $\eta < \frac{1}{2}$ and $\beta = O(\tau(1/2-\eta)) = \tau/\Omega(\frac{1}{1/2-\eta})$. This classification noise tolerance matches the information theoretic limit of $1/2$ and is therefore optimal. We next demonstrate the optimality of the dependence of β on η and τ in the following two theorems. Together, these theorems show the error tolerance of Theorem 3 to be within a $\log(1/\tau)$ factor of optimal.

Theorem 5 *There is no class which is PAC learnable with CAM* $\forall\eta < 1/2$ *and* $\beta = g_1(\tau)/o(\frac{1}{1/2-\eta})$ *for any function* $g_1(\cdot)$.

Proof: If $\beta = g_1(\tau)/o(\frac{1}{1/2-\eta})$, then for any fixed τ (and therefore fixed $g(\tau)$) there exists an $\eta < 1/2$ for which $\beta \geq g_1(\tau)/[\frac{g_1(\tau)}{1/2-\eta}] = 1/2-\eta$, in which case $\eta+\beta \geq 1/2$. But no class is learnable under such conditions since the malicious adversary is then able to make the CAM oracle simulate a standard classification noise oracle with noise rate $1/2$, making every label look like a random bit. □

Recall that Kearns and Li [12] prove an upper bound of $\frac{\varepsilon}{1+\varepsilon}$ on the amount of malicious error tolerable when learning any *distinct* concept class. We use this result in the next theorem.

Theorem 6 *No distinct class is PAC learnable with CAM of* $\beta = \omega(\tau\log(1/\tau))\cdot g_2(1/2-\eta)$ *for any function* $g_2(\cdot)$.

Proof: Aslam and Decatur [4] have shown that if a class is SQ learnable, then it is SQ learnable with $\tau \geq \frac{\varepsilon}{poly(n)\log(1/\varepsilon)}$. Therefore, for fixed n and fixed $\eta < \frac{1}{2}$, $\beta = \omega(\tau\log(1/\tau))\cdot g_2(1/2-\eta) = \omega(\varepsilon)$. However, as stated above, no distinct class is learnable with $\beta \geq \frac{\varepsilon}{1+\varepsilon} = \Theta(\varepsilon)$. □

4 Other Hybrid Noise Models

In Section 3.1 we defined a hybrid noise environment in which an example oracle exhibited both classification noise and malicious errors. We may instead define hybrid noise environments based on other types of noise models. Two other such "noise" models are *distribution shift* and *distribution restricted* learning [8]. In both models, the examples are labelled correctly, but the distribution of examples is somehow modified. We show that any combination of these noise models, along with classification noise and/or malicious errors, define a hybrid noise environment in which we can efficiently simulate any statistical query algorithm. Once again, the statistical query algorithm quantifies how much error we can tolerate in estimating a query. To each individual type of noise, we may allocate a part of that fault tolerance. Thus the tolerance of the SQ algorithm quantifies the generic fault tolerance of the algorithm.

Another result on learning in a hybrid noise environment incorporates both attribute and classification noise. Specifically, the technique of simulating SQ algorithms to achieve learnability in hybrid noise models has also been exploited by Decatur and Gennaro [9] in order to characterize learnability in a model combining attribute noise and classification noise. Here again, the SQ algorithm characterizes the amount of noise tolerance that may be achieved.

4.1 Individual Noise Models

In the model of distribution shift, the learning algorithm is given access to an example oracle which draws examples according to a distribution which differs from the distribution on which it will be tested. Recall that examples drawn are always labelled correctly. The only restriction on the distribution used for training is that it be "close" to the distribution used for testing. A class is learnable with distribution shift σ if there exists an algorithm which can tolerate training on any distribution which is within distance σ from the testing distribution. The distance between two distributions P and Q over the domain X is defined as follows:

$$d(P,Q) = \max_{A \subseteq X} |P(A) - Q(A)|.$$

Note that this definition of distance is equivalent to $d(P,Q) = \frac{1}{2}\sum_{x \in X} |P(x) - Q(x)|$. If $\mathcal{F}$ is PAC learnable, then it is learnable with distribution shift σ if and only if $\sigma = O(\varepsilon)$ [8].

In distribution restricted learning, a learner is promised that the distribution of examples belongs to a specified class of distributions $\mathcal{D}$. We can use a distribution restricted SQ algorithm to learn on distributions outside of the promised class of distributions $\mathcal{D}$. The larger class of distributions on which learning is possible is determined by the tolerance τ of the SQ algorithm. Specifically, the larger class of distributions is composed of distributions within distance $O(\tau)$ of some distribution in $\mathcal{D}$ [8].

4.2 The Hybrid Noise Model

The model which combines all four of these individual noise models is described by the following learning environment. There is a testing distribution D_1 which is close to some distribution $D \in \mathcal{D}$ (*i.e.* , $d(D, D_1) \leq \gamma$) and there is a training distribution D_2 which is close to D_1 (*i.e.* , $d(D_1, D_2) \leq \sigma$). The example oracle used for training is as follows:

$$EX_{\text{HYB}}^{\eta,\beta,\gamma,\sigma}(f, D_1) = \begin{cases} EX(f, D_2) & 1 - \eta - \beta \\ EX(\neg f, D_2) & \eta \\ \text{Adversary} & \beta \end{cases}$$

An SQ algorithm for the class $\mathcal{F}$ on distributions $\mathcal{D}$ can be converted (by simulating statistical queries) into a PAC algorithm for learning $\mathcal{F}$ in this environment. The simulation is valid for all $\eta < 1/2$, and $\beta + \gamma + \sigma = O(\tau(1/2 - \eta))$. One again, there is a smooth tradeoff between all of the different types of noise and the amount of noise tolerable can be shown roughly optimal in terms of its dependence on τ. The simulations and proofs of optimality are similar to those show earlier and details are given in the full paper.

Acknowledgements

Thanks to Robert Schapire whose questions prompted this research and to Javed Aslam and Leslie Valiant for helpful comments on earlier drafts of this paper.

References

[1] Dana Angluin. Computational learning theory: Survey and selected bibliography. In *Proceedings of the* 24^{th} *Annual ACM Symposium on the Theory of Computing*, 1992.

[2] Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

[3] Dana Angluin and Leslie G. Valiant. Fast probabilistic algorithms for Hamiltonian circuits and matchings. *Journal of Computer and System Sciences*, 18(2):155–193, April 1979.

[4] Javed Aslam and Scott Decatur. General bounds on statistical query learning and PAC learning with noise via hypothesis boosting. In *Proceedings of the* 34^{th} *Annual Symposium on Foundations of Computer Science*, pages 282–291, November 1993.

[5] Javed Aslam and Scott Decatur. A general lower bound on the sample complexity of PAC learning in the presence of classification noise. In preparation, 1994.

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

[7] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–865, 1989.

[8] 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.

[9] Scott Decatur and Rosario Gennaro. Learning over a noisy communication channel. In preparation, 1994.

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

[11] 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.

[12] Michael Kearns and Ming Li. Learning in the presence of malicious errors. In *Proceedings of the* 20^{th} *Annual ACM Symposium on Theory of Computing*, Chicago, Illinois, May 1988.

[13] Philip D. Laird. *Learning from Good and Bad Data.* Kluwer international series in engineering and computer science. Kluwer Academic Publishers, Boston, 1988.

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

[15] Leslie Valiant. Learning disjunctions of conjunctions. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, 1985.