On the Sample Complexity of Noise-Tolerant Learning

Javed A. Aslam*
Department of Computer Science
Dartmouth College
Hanover, NH 03755

Scott E. Decatur†
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

In this paper, we further characterize the complexity of noise-tolerant learning in the PAC model. Specifically, we show a general lower bound of $\Omega\left(\frac{\log(1/\delta)}{\varepsilon(1-2\eta)^2}\right)$ on the number of examples required for PAC learning in the presence of classification noise. Combined with a result of Simon, we effectively show that the sample complexity of PAC learning in the presence of classification noise is $\Omega\left(\frac{\mathrm{VC}(\mathcal{F})}{\varepsilon(1-2\eta)^2} + \frac{\log(1/\delta)}{\varepsilon(1-2\eta)^2}\right)$. Furthermore, we demonstrate the optimality of the general lower bound by providing a noise-tolerant learning algorithm for the class of symmetric Boolean functions which uses a sample size within a constant factor of this bound. Finally, we note that our general lower bound compares favorably with various general upper bounds for PAC learning in the presence of classification noise.

Keywords

Machine Learning, Computational Learning Theory, Computational Complexity, Fault Tolerance, Theory of Computation

1 Introduction

In this paper, we derive bounds on the complexity of learning in the presence of noise. We consider the "Probably Approximately Correct" (PAC) model of learning introduced by Valiant [11]. In this setting, a learner is given the task of determining a close approximation of an unknown $\{0, 1\}$-valued target function f. The learner is given $\mathcal{F}$, a class of functions to which f belongs, and accuracy and confidence parameters ε and δ. The learner gains information about the target function by viewing examples which are labelled according to f. The learner is required to output an hypothesis such that, with high confidence (at least $1 - \delta$), the accuracy of the hypothesis is high (at least $1 - \varepsilon$). Two standard complexity measures studied in the PAC model are *sample complexity*, the number of examples used or required by a PAC learning algorithm, and *time complexity*, the computation time used or required by a PAC learning algorithm.

*This work was performed while the author was at Harvard University and supported by Air Force Contract F49620-92-J-0466. Author's current net address: `jaa@cs.dartmouth.edu`

†This work was performed while the author was at Harvard University and supported by an NDSEG Doctoral Fellowship and by NSF Grant CCR-92-00884. Author's current net address: `sed@theory.lcs.mit.edu`

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." A widely studied model of noise for both theoretical and experimental research is the *classification noise* model introduced by Angluin and Laird [1]. In this model, each example received by the learner is mislabelled randomly and independently with some fixed probability $\eta < 1/2$.

It is not surprising that algorithms for learning in the presence of classification noise use more examples than their corresponding noise-free algorithms. It is therefore natural to ask, "What is the increase in the complexity of learning when the data used for learning is corrupted by noise?" We focus on the number of examples needed for learning in the presence of classification noise.

Previous attempts at lower bounds on the sample complexity of classification noise learning yielded suboptimal results and in some cases relied on placing restrictions on the learning algorithm. Laird [8] showed that a specific learning algorithm, one which simply chooses the function in the target class $\mathcal{F}$ with the fewest disagreements on the sample of data, requires $\Omega(\frac{\log(|\mathcal{F}|/\delta)}{\varepsilon(1-2\eta)})$ examples. Note that this result is only applicable to *finite* target classes. Simon [9] showed that any algorithm for learning in the presence of classification noise requires $\Omega(\frac{\mathrm{VC}(\mathcal{F})}{\varepsilon(1-2\eta)^2})$ examples where $\mathrm{VC}(\mathcal{F})$ is the Vapnik-Chervonenkis dimension of $\mathcal{F}$.[1] One could also consider a general lower bound on the sample complexity of noise-free learning to be a general lower bound on the sample complexity of classification noise learning. The noise-free bound of Ehrenfeucht *et al.* [6] and Blumer *et al.* [4] states that $\Omega(\frac{\mathrm{VC}(\mathcal{F})}{\varepsilon} + \frac{\log(1/\delta)}{\varepsilon})$ examples are required for learning.

In this paper, we show a general lower bound of $\Omega(\frac{\log(1/\delta)}{\varepsilon(1-2\eta)^2})$ on the number of examples required for PAC learning in the presence of classification noise. Combined with the above result of Simon, we effectively show that the sample complexity of PAC learning in the presence of classification noise is $\Omega(\frac{\mathrm{VC}(\mathcal{F})}{\varepsilon(1-2\eta)^2} + \frac{\log(1/\delta)}{\varepsilon(1-2\eta)^2})$. In addition to subsuming previous lower bounds, this result is completely general in that it holds for any algorithm which learns in the presence of classification noise, regardless of the amount of computation time allowed and regardless of how expressive an hypothesis class is used (including general randomized prediction hypotheses). Note that the above bound is a generalized analog of the noise-free bound of Ehrenfeucht *et al.* and Blumer *et al.*

We demonstrate the asymptotic optimality of the combined general lower bound by showing a specific upper bound for learning symmetric functions over the Boolean hypercube $\{0,1\}^n$ using a sample size which is within a constant factor of the general lower bound. The learning algorithm we give uses the optimal sample complexity, runs in polynomial time, and outputs a deterministic hypothesis from the target class. We therefore demonstrate that not only is the general lower bound optimal when placing no restrictions on the learning algorithm, but it cannot be improved even if

[1] $\mathrm{VC}(\mathcal{F})$ is a combinatorial characterization of $\mathcal{F}$ which usually depends on n, the common length of the elements from the domain of functions in $\mathcal{F}$.

one were to restrict the learning algorithm to run in polynomial time and to require it to output an hypothesis from the target class $\mathcal{F}$, the most restrictive possible hypothesis class.

We finally note that our general lower bound is quite close to the various fairly general upper bounds known to exist. Laird [8] has shown that for finite classes, a sample of size $O(\frac{\log(|\mathcal{F}|/\delta)}{\varepsilon(1-2\eta)^2})$ is sufficient for classification noise learning. This result is not computationally efficient in general, since it relies on the ability to minimize disagreements with respect to a sample. The results of Talagrand [10] imply that for classes of finite VC-dimension, a sample of size $O(\frac{\mathrm{VC}(\mathcal{F})}{\varepsilon^2(1-2\eta)^2} + \frac{1}{\varepsilon^2(1-2\eta)^2}\log\frac{1}{\varepsilon(1-2\eta)\delta})$ is sufficient for classification noise learning. This result also relies on the ability to minimize disagreements. Finally, Aslam and Decatur [3] have shown that a sample of size $\tilde{O}(\frac{poly(n)+\log(1/\delta)}{\varepsilon^2(1-2\eta)^2})$ is sufficient for *polynomial time* classification noise learning of any class known to be learnable in the statistical query model.[2]

2 Definitions

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

In an instance of PAC learning, a learner is given the task of determining a close approximation of an unknown $\{0,1\}$-valued target function from labelled examples of that function. The unknown target function f is assumed to be an element of a known function class $\mathcal{F}$ defined over an instance space X. The instance space X is typically either the Boolean hypercube $\{0,1\}^n$ or n-dimensional Euclidean space $\Re^n$. We use the parameter n to denote the common length of instances $x \in X$.

We assume that the instances are distributed according to some unknown probability distribution D on X. The learner is given access to an example oracle $EX(f, D)$ as its source of data. A call to $EX(f, D)$ returns a labelled example $\langle x, l\rangle$ where the instance $x \in X$ is drawn randomly and independently according to the unknown distribution D, and the label $l = f(x)$. We often refer to a sequence of labelled examples drawn from an example oracle as a *sample*.

A learning algorithm draws a sample from $EX(f, D)$ and eventually outputs an hypothesis h. For any hypothesis h, the *error rate* of h is defined to be the probability that $h(x) \neq f(x)$ for an instance $x \in X$ drawn randomly according to D. Although we often allow the learning algorithm to output any hypothesis it chooses (including general, possibly randomized, programs), in some cases we consider the complexity of learning algorithms which are required to output an hypothesis from a specific *representation class* $\mathcal{H}$.

The learner's goal is to output, with probability at least $1-\delta$, an hypothesis h whose error rate is at most ε, for the given *error parameter* ε and *confidence parameter* δ. A learning algorithm is said to be *polynomially efficient* if its running time is polynomial in $1/\varepsilon$, $1/\delta$, and n.

[2] $\tilde{O}$ denotes an asymptotic upper bound ignoring lower order, typically logarithmic, factors. The variable n parameterizes the complexity of the target class as described in the following section.

In the classification noise variant of PAC learning, the learning algorithm no longer has access to $EX(f,D)$, but instead has access 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 classification $f(x)$, but independently returns $\langle x, f(x)\rangle$ with probability $1-\eta$ or $\langle x, \neg f(x)\rangle$ with probability η. The learner is allowed to run in time polynomial in $1/(1-2\eta)$ and the standard parameters,[3] but is still required to output an hypothesis which is ε-good with respect to *noise-free* data.

Finally, we characterize a concept class $\mathcal{F}$ by its Vapnik-Chervonenkis dimension, VC($\mathcal{F}$), defined as follows. For any concept class $\mathcal{F}$ and set of instances $S = \{x_1, \ldots, x_d\}$, we say that $\mathcal{F}$ shatters S if for all of the 2^d possible binary labellings of the instances in S, there is a function in $\mathcal{F}$ that agrees with that labelling. VC($\mathcal{F}$) is the cardinality of the largest set shattered by $\mathcal{F}$.

3 The General Lower Bound

In this section, we prove the following general lower bound on the sample complexity required for PAC learning in the presence of classification noise:

Theorem 1 *For all classes $\mathcal{F}$ such that* $\mathrm{VC}(\mathcal{F}) \geq 2$, *and for all $\varepsilon \leq 1/3$, $\delta < 1/20$ and $\eta \geq 29/60$, PAC learning $\mathcal{F}$ in the presence of classification noise requires a sample of size $m = \frac{1}{100\varepsilon(1-2\eta)^2} \ln \frac{1}{5\delta}$.*

Proof: We begin by noting that if $\mathcal{F}$ has VC-dimension at least 2, then there must exist two instances which can be labelled in all possible ways by functions in $\mathcal{F}$. Let x and y be such instances, and let f_0 and f_1 be functions in $\mathcal{F}$ which label x and y as follows: $f_0(x) = f_1(x) = 0$, $f_0(y) = 0$, and $f_1(y) = 1$.

By the definition of PAC learning in the presence of classification noise, any valid algorithm for learning a function class $\mathcal{F}$ in the presence of noise must output an accurate hypothesis, with high probability, given a noisy example oracle corresponding to *any* target function and *any* distribution over the instances. Thus, as an adversary, we may choose both the distribution over the instances and the target functions of interest. Let the distribution D be defined by $D(y) = 3\varepsilon$ and $D(x) = 1 - 3\varepsilon$, and consider the functions f_0 and f_1.

First, note that with respect to the distribution chosen, the functions f_0 and f_1 are fairly dissimilar. Each function has an error rate of 3ε with respect to the other, and therefore no hypothesis can be ε-good with respect to both functions. In some sense, the learning algorithm must "decide" whether the instance y should be labelled 1 or 0 given the data that it receives ¿from the noisy example oracle. However, given a "small" sample containing "relatively few" y-instances, it may not be possible to confidently make this determination in the presence of noise. This is essentially the idea behind the proof that follows.

[3] If the learner only has access to η_b, an upper bound on the noise rate η, then it is allowed to run in time polynomial in $1/(1-2\eta_b)$.

Let $m = \frac{1}{100\varepsilon(1-2\eta)^2} \ln \frac{1}{5\delta}$ be the sample size in question, and let $S = (X \times \{0,1\})^m$ be the set of all samples of size m.[4] For a sample $s \in S$, let $P_i(s)$ denote the probability of drawing sample s from $EX^\eta{}_{\mathrm{CN}}(f_i, D)$. If A is a learning algorithm for $\mathcal{F}$, let $A_i(s)$ be the probability that on input s, algorithm A outputs an hypothesis which has error rate greater than ε for f_i with respect to D. Let F_i be the probability that algorithm A *fails*, *i.e.* outputs an hypothesis with error rate greater than ε, given a randomly drawn sample of size m from $EX^\eta{}_{\mathrm{CN}}(f_i, D)$. We then have

$$F_i = \sum_S A_i(s) P_i(s).$$

Thus, if A learns $\mathcal{F}$, it must be the case that both $F_0 \leq \delta$ and $F_1 \leq \delta$. We show that for the sample size m given above, both of these conditions cannot hold, and therefore A does not learn $\mathcal{F}$.

We assume without loss of generality that $F_0 \leq \delta$ and show that $F_1 > \delta$. For a given sample s, let $g_1(s)$ be the fraction of the examples in s which are y-instances. Furthermore, let $g_2(s)$ be the fraction of the y-instances in s which are labelled 1. We define the following subsets of the sample space S:

$$\begin{aligned} S_1 &= \{s \in S : g_1(s) \in [2\varepsilon, 4\varepsilon]\} \\ S_2 &= \{s \in S_1 : g_2(s) \in [\eta - 5(1-2\eta), (1-\eta) + 5(1-2\eta)]\} \end{aligned}$$

The set of samples S_2 contains "likely" samples, regardless of which function is the target. Note that since $F_0 \leq \delta$, we clearly have

$$\sum_{S_2} A_0(s) P_0(s) \leq \delta.$$

Similarly, we have

$$F_1 \geq \sum_{S_2} A_1(s) P_1(s).$$

It is this last summation which we show to be greater than δ.

Note that by the construction of the distribution D, any hypothesis which is ε-good with respect to f_0 must be ε-bad with respect to f_1, and vice versa. Thus, for any sample s,

$$A_0(s) + A_1(s) \geq 1.$$

We therefore have the following:

$$F_1 \geq \sum_{S_2} A_1(s) P_1(s) \quad \geq \quad \sum_{S_2} (1 - A_0(s)) P_1(s)$$

[4]Note that it will be shown that a sample of size m is insufficient for learning, and this clearly implies that a sample of size *at most* m is also insufficient for learning.

$$= \sum_{S_2} P_1(s) - \sum_{S_2} A_0(s) P_1(s) \tag{1}$$

In order to lower bound F_1, in Lemma 2 we lower bound the first summation on the right-hand side of Equation 1, and in Lemma 3 we upper bound the second summation. We make use of the following bounds on the tail of the binomial distribution [2, 5, 7]:

Lemma 1 *For $p \in [0,1]$ and positive integer m, let $LE(p,m,r)$ denote the probability of at most an r fraction of successes in m independent trials of a Bernoulli random variable with probability of success p. Let $GE(p,m,r)$ denote the probability of at least an r fraction of successes. Then for $\alpha \in [0,1]$,*

$$\begin{aligned}
LE(p,m,(1-\alpha)p) &\leq e^{-\alpha^2 mp/2} \\
GE(p,m,(1+\alpha)p) &\leq e^{-\alpha^2 mp/3} \\
LE(p,m,(p-\alpha)) &\leq e^{-2\alpha^2 m} \\
GE(p,m,(p+\alpha)) &\leq e^{-2\alpha^2 m}.
\end{aligned}$$

Lemma 2 $\sum\limits_{S_2} P_1(s) > 1/4$.

Proof: In order to lower bound $\sum_{S_2} P_1(s)$, we first lower bound $\sum_{S_1} P_1(s)$. The probability that a sample has more than $4\varepsilon m$ y-instances or fewer than $2\varepsilon m$ y-instances is upper bounded by Equations 2 and 3, respectively.

$$GE(3\varepsilon, m, 3\varepsilon(1+1/3)) \leq e^{-m3\varepsilon(1/3)^2/3} = e^{-\varepsilon m/9} \tag{2}$$

$$LE(3\varepsilon, m, 3\varepsilon(1-1/3)) \leq e^{-m3\varepsilon(1/3)^2/2} = e^{-\varepsilon m/6} \tag{3}$$

Given $m = \frac{1}{100\varepsilon(1-2\eta)^2} \ln \frac{1}{5\delta}$, $\delta < 1/20$ and $\eta \geq 29/60$, we have $m > \frac{9}{\varepsilon} \ln 4$. Thus, the probabilities in Equations 2 and 3 are each less than $1/4$. Therefore, with probability greater than $1/2$, a sample of size m drawn randomly from $EX^{\eta}{}_{\mathrm{CN}}(f_1, D)$ is an element of S_1, *i.e.* $\sum_{S_1} P_1(s) > 1/2$. We next determine the probability of a sample being in S_2 given that it is in S_1 and the target function is f_1. Thus, we may assume that the fraction of y-instances is at least 2ε. Given that a sample has at least a 2ε fraction of y-instances, the probability that the fraction of its y-instances labelled 1 is more than a $(1-\eta) + 5(1-2\eta)$ or less than $\eta - 5(1-2\eta)$, is upper bounded by Equations 4 and 5, respectively.

$$GE(1-\eta, 2\varepsilon m, (1-\eta) + 5(1-2\eta)) \leq e^{-4\varepsilon m(5(1-2\eta))^2} = e^{-100\varepsilon m(1-2\eta)^2} \tag{4}$$

$$LE(1-\eta, 2\varepsilon m, (1-\eta) - 6(1-2\eta)) \leq e^{-4\varepsilon m(6(1-2\eta))^2} = e^{-144\varepsilon m(1-2\eta)^2} \tag{5}$$

Given $m = \frac{1}{100\varepsilon(1-2\eta)^2} \ln \frac{1}{5\delta}$ and $\delta < 1/20$, we have $m > \frac{1}{100\varepsilon(1-2\eta)^2} \ln 4$. Thus, the probabilities in Equations 4 and 5 are each less than $1/4$. Therefore, given that the sample drawn is in S_1, with probability greater than 1/2, the sample drawn is in S_2, *i.e.* $\sum_{S_2} P_1(s|S_1) > 1/2$. Since $P_1(s|S_1) = P_1(s)/\sum_{s' \in S_1} P_1(s')$, we have $\sum_{S_2} P_1(s) = [\sum_{S_2} P_1(s|S_1)] \cdot [\sum_{S_1} P_1(s)] > 1/2 \cdot 1/2 = 1/4$. ■

Lemma 3 *If* $\sum\limits_{S_2} A_0(s)P_0(s) \leq \delta$*, then* $\sum\limits_{S_2} A_0(s)P_1(s) < 1/5$*.*

Proof: Let T^i be the set of samples in S_2 with i y-instances and $m-i$ x-instances. Let $T^i_{j,k} \subset T^i$ be the set of samples in S_2 with i y-instances and $m-i$ x-instances, where j of the y-instances are labelled 1 and k of the x-instances are labelled 1. For any $s \in T^i_{j,k}$, we have

$$\begin{aligned} P_0(s) &= (3\varepsilon)^i(1-3\varepsilon)^{m-i} \cdot \eta^j(1-\eta)^{i-j} \cdot \eta^k(1-\eta)^{m-i-k} \\ P_1(s) &= (3\varepsilon)^i(1-3\varepsilon)^{m-i} \cdot (1-\eta)^j\eta^{i-j} \cdot \eta^k(1-\eta)^{m-i-k}. \end{aligned}$$

Therefore, for any $s \in T^i_{j,k}$, we have

$$P_1(s) = P_0(s)\left(\frac{1-\eta}{\eta}\right)^{2j-i}.$$

Let $v_1 = 2\varepsilon$, $v_2 = 4\varepsilon$, $v_3 = \eta - 5(1-2\eta)$, and $v_4 = (1-\eta) + 5(1-2\eta)$. We may now rewrite the desired summation as follows:

$$\begin{aligned} \sum_{S_2} A_0(s)P_1(s) &= \sum_{i=v_1 m}^{v_2 m} \sum_{j=v_3 i}^{v_4 i} \sum_{k=0}^{m-i} \sum_{T^i_{j,k}} A_0(s)P_1(s) \\ &= \sum_{i=v_1 m}^{v_2 m} \sum_{j=v_3 i}^{v_4 i} \sum_{k=0}^{m-i} \sum_{T^i_{j,k}} A_0(s)P_0(s)\left(\tfrac{1-\eta}{\eta}\right)^{2j-i}. \end{aligned}$$

Note that $2j - i \leq 2v_4 i - i = (2v_4 - 1)i \leq (2v_4 - 1)v_2 m$. We therefore obtain:

$$\begin{aligned} \sum_{S_2} A_0(s)P_1(s) &\leq \left(\tfrac{1-\eta}{\eta}\right)^{v_2 m(2v_4-1)} \sum_{i=v_1 m}^{v_2 m} \sum_{j=v_3 i}^{v_4 i} \sum_{k=0}^{m-i} \sum_{T^i_{j,k}} A_0(s)P_0(s) \\ &= \left(\tfrac{1-\eta}{\eta}\right)^{v_2 m(2v_4-1)} \sum_{S_2} A_0(s)P_0(s) \\ &\leq \delta\left(\tfrac{1-\eta}{\eta}\right)^{v_2 m(2v_4-1)}. \end{aligned} \tag{6}$$

Using the fact that for all z, $1 + z \leq e^z$, we have $\frac{1-\eta}{\eta} = 1 + \frac{1-2\eta}{\eta} \leq e^{(1-2\eta)/\eta}$. Given the expression $m = \frac{1}{100\varepsilon(1-2\eta)^2} \ln \frac{1}{5\delta}$ and the constraint $\eta \geq 29/60 > 44/100$, we then have

$$\begin{aligned} \left(\tfrac{1-\eta}{\eta}\right)^{v_2 m(2v_4-1)} &\leq e^{(1-2\eta)v_2 m(2v_4-1)/\eta} \\ &= e^{\ln(1/5\delta)\cdot 44/100\eta} < 1/5\delta. \end{aligned} \tag{7}$$

By Equations 6 and 7, we have $\sum_{S_2} A_0(s)P_1(s) < 1/5$. ■

Combining Equation 1 with Lemmas 2 and 3, we have $F_1 > 1/20 > \delta$. Therefore, if on a sample of size m, the algorithm fails with probability at most δ when the target is f_0, then the algorithm must fail with probability more than δ when the target is f_1. □

Note that we have not attempted to optimize the constants in this lower bound.

3.1 The Combined Lower Bound

Simon [9] proves the following lower bound:

Theorem 2 *PAC learning a function class $\mathcal{F}$ in the presence of classification noise requires a sample of size* $\Omega(\frac{\mathrm{VC}(\mathcal{F})}{\varepsilon(1-2\eta)^2})$.

Combining Theorems 1 and 2 we obtain the following lower bound on the number of examples required for PAC learning in the presence of classification noise.

Theorem 3 *PAC learning a function class $\mathcal{F}$ in the presence of classification noise requires a sample of size* $\Omega(\frac{\mathrm{VC}(\mathcal{F})}{\varepsilon(1-2\eta)^2} + \frac{\log(1/\delta)}{\varepsilon(1-2\eta)^2})$.

Note that the result obtained is general in the sense that it holds for all PAC learning problems. It holds for any algorithm, whether deterministic or randomized, and any hypothesis representation class, even general (possibly probabilistic) prediction programs. Furthermore, the result holds for all algorithms independent of the computational resources used. Finally, note that Theorem 3 is a generalized analog of the general lower bound of Blumer *et al.* [4] and Ehrenfeucht *et al.* [6] for noise-free learning.

4 Optimality of the General Lower Bound

In this section, we show that the general lower bound of Theorem 3 is asymptotically optimal in a very strong sense. First consider the upper bound for learning finite classes in the presence of classification noise due to Laird [8]. Laird has shown that a sample of size $O(\frac{\log(|\mathcal{F}|/\delta)}{\varepsilon(1-2\eta)^2})$ is sufficient for classification noise learning. Since many finite classes have the property that $\log|\mathcal{F}| = \Theta(\mathrm{VC}(\mathcal{F}))$, this result implies that a sample of size $O(\frac{\mathrm{VC}(\mathcal{F})}{\varepsilon(1-2\eta)^2} + \frac{\log(1/\delta)}{\varepsilon(1-2\eta)^2})$. is sufficient for learning these classes in the presence of classification noise. Thus, from an information-theoretic standpoint, the lower bound of Theorem 3 is asymptotically optimal. However, the upper bound of Laird is non-computational in general since it relies on the ability to minimize disagreements with respect to a sample, a problem known to be NP-complete for many classes. One might imagine that a better general lower bound may exist if one were to restrict learning algorithms to run in polynomial time. In the theorem below, we show that such a lower bound cannot exist. We do so by giving an algorithm for learning the class of symmetric Boolean functions in the presence of classification noise whose sample complexity matches that of Theorem 3. Furthermore, this algorithm runs in polynomial time and outputs an hypothesis from the class of target functions. Thus, the general lower bound cannot be improved even if one were to restrict learning algorithms to work in polynomial time and output hypotheses from the most restrictive representation class, the target class.

The class $\mathcal{S}$ of symmetric functions over the domain $\{0,1\}^n$ is the set of all Boolean functions f for which $H(x) = H(y)$ implies $f(x) = f(y)$ where $H(x)$ represents the number of components which are 1 in the Boolean vector x. Note that there are 2^{n+1} functions in $\mathcal{S}$ and that the VC-dimension of $\mathcal{S}$ is $n+1$.

Theorem 4 *The class $\mathcal{S}$ of symmetric functions is learnable by symmetric functions in polynomial time by an algorithm which uses a sample of size $O(\frac{\mathrm{VC}(\mathcal{S})}{\varepsilon(1-2\eta)^2} + \frac{\log(1/\delta)}{\varepsilon(1-2\eta)^2})$.*

Proof: We make use of a result of Laird [8] on learning a class $\mathcal{F}$ in the presence of classification noise which states that if an algorithm uses a sample of size $\frac{8\ln(|\mathcal{F}|/\delta)}{3\varepsilon(1-2\eta)^2}$ and outputs the function in the class $\mathcal{F}$ which has the fewest disagreements with the sample, then this hypothesis is ε-good with probability at least $1-\delta$. By the size and VC-dimension of $\mathcal{S}$ described above, the sample complexity stated in the theorem is achieved. To complete the proof, we design an algorithm which, given a sample, outputs in polynomial time the symmetric function with the fewest disagreements on that sample.

For each value $i \in \{0, \ldots, n\}$, our algorithm computes the number of labelled examples with i 1's in the instance which are labelled positive and the number of such examples labelled negative. If the number of positive examples is more than the number of negative examples, then the hypothesis we construct outputs 1 on all instances with i 1's, otherwise it outputs 0 on such examples. Clearly no other symmetric function has fewer disagreements with the sample. □

References

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

[2] 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.

[3] 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.

[4] 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.

[5] Herman Chernoff. A measure of the asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.*, 23:493–507, 1952.

[6] Andrzej Ehrenfeucht, David Haussler, Michael Kearns, and Leslie Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–251, September 1989.

[7] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

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

[9] Hans Ulrich Simon. General bounds on the number of examples needed for learning probabilistic concepts. In *Proceedings of the Sixth Annual ACM Workshop on Computational Learning Theory*, pages 402–411. ACM Press, 1993.

[10] M. Talagrand. Sharper bounds for empirical processes. To appear in *Annals of Probability and Its Applications*.

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