Uniform-Distribution Attribute Noise Learnability

Nader H. Bshouty
Department of Computer Science, Technion, Haifa, Israel
Email: bshouty@cs.technion.ac.il

Jeffrey C. Jackson
Department of Mathematics and Computer Science, Duquesne University, Pittsburgh, U.S.A.
Email: jackson@mathcs.duq.edu

Christino Tamon*
Department of Mathematics and Computer Science, Clarkson University, Potsdam, U.S.A.
Email: tino@clarkson.edu

May 23, 2003

Abstract

We study the problem of PAC-learning Boolean functions with random attribute noise under the uniform distribution. We define a *noisy distance* measure for function classes and show that if this measure is small for a class $\mathcal{C}$ and an attribute noise distribution D then $\mathcal{C}$ is not learnable with respect to the uniform distribution in the presence of noise generated according to D. The noisy distance measure is then characterized in terms of Fourier properties of the function class. We use this characterization to show that the class of all parity functions is not learnable for any but very concentrated noise distributions D. On the other hand, we show that if $\mathcal{C}$ is learnable with respect to uniform using a standard Fourier-based learning technique, then $\mathcal{C}$ is learnable with time and sample complexity also determined by the noisy distance. In fact, we show that this style algorithm is nearly the best possible for learning in the presence of attribute noise. As an application of our results, we show how to extend such an algorithm for learning AC^0 so that it handles certain types of attribute noise with relatively little impact on the running time.
Keywords: computational learning theory; learning with noise; Fourier analysis.

1 Introduction

The problem of attribute noise in PAC-learning was studied originally by Shackelford and Volper [8] for the case of k-DNF expressions. Their *uniform* attribute noise model consists of a Bernoulli process that will either flip or not flip each attribute value with a fixed probability $p \in [0, 1]$ that is the same for every attribute. While Shackelford and Volper assumed that the learner knows the noise rate p, Goldman and Sloan [4] proved that this assumption is not necessary in order to learn monomials.

In addition to uniform attribute noise, Goldman and Sloan also considered a *product* noise model in which there are n noise rates p_i, one for each distinct attribute x_i, $i \in [n]$. They showed that if the product noise rates p_i are unknown, then no PAC-learning algorithm exists that can

*Corresponding author

tolerate a noise rate higher than 2ϵ, where ϵ is the required-accuracy parameter for PAC learning. Subsequently, Decatur and Gennaro [3] proved that if the different noise rates are *known* (or if some upper bound on them is given) then there exist efficient PAC-learning algorithms for simple classes such as monomials and k-DNF expressions.

In this paper we consider a very general attribute noise model, but limit the distribution that will be used to generate examples and to evaluate the accuracy of the hypothesis generated by the learning algorithm. Specifically, we focus on the problem of PAC learning with respect to the uniform distribution over examples, with little or no constraint on the distribution used to generate attribute noise in the examples. We give both lower and upper bounds.

First, we define a measure of *noisy distance* for concept classes and show that the sample size required for PAC learning a class over the uniform distribution is inversely proportional to the noisy distance of the class. We also give a characterization of the noisy distance in terms of Fourier properties of the class. As an example of how this characterization can be used, we show that the class of all parity functions is not (even information theoretically) PAC learnable with respect to uniform unless the attribute noise distribution puts nonnegligible weight on one or more of the bit-vectors representing the noise to be applied to an example. So, for example, if the attribute noise is applied by independently flipping a coin with constant bias for each bit then the maximum weight on any noise vector is exponentially small, implying that the parity class is not learnable for this noise distribution. This holds even if the noise process is known. On the other hand, we observe as a corollary of a result of Blum, Burch, and Langford [1] that the class of monotone Boolean functions is weakly PAC-learnable even if the noise process is unknown.

We then turn to developing positive learnability results. Specifically, we show that any concept class that is PAC-learnable with respect to the uniform distribution using an algorithm in the style of Linial, Mansour, and Nisan [7] can be adapted to handle attribute noise, assuming the probability distribution of the noise process is known. However, the noisy distance of a class depends on the noise distribution, so the sample complexity of our algorithm is dependent on the noise process as well as the usual PAC factors. The dependence of the the sample complexity of our algorithm matches, to within polynomial factors, our lower bound for learning with attribute noise. We then apply our theory to show that for a specific class of noise distributions—mild known rates of attribute noise independently applied to the inputs—AC^0 remains learnable with respect to the uniform distribution in time comparable to that of the best known noise-free bound.

Our Fourier techniques share some commonalities with methods developed by Benjamini, Kalai, and Schramm [2] in their work that studied percolation and its relation to noise sensitivity of Boolean functions. Their techniques, like ours, were strongly motivated by the influential work of Kahn, Kalai, and Linial [6] on Fourier analysis of Boolean functions.

2 Definitions and Notation

The problem considered in this paper is *PAC learning* Boolean functions under some fixed distribution over instances when *attribute noise* is also applied to the instances. To a lesser extent, we also consider *classification noise*. We define these concepts more precisely below. For simplicity, our definitions suppress some standard details (particularly the notion of *size* of functions) that are not critical to the results in this paper.

For a natural number n, we consider classes of Boolean functions $f : \{0,1\}^n \to \{-1,+1\}$ and distributions over $\{0,1\}^n$. The uniform distribution on $\{0,1\}^n$ is denoted U_n (or just U when n is understood from context), i.e., $U_n(x) = 2^{-n}$, for all $x \in \{0,1\}^n$. The *bitwise exclusive-or* of two n-bit vectors $a, b \in \{0,1\}^n$ is denoted $a \oplus b$. The *unit vector* $e_i \in \{0,1\}^n$ has its i-th bit set to one and

all other bits set to zero. For $a \in \{0,1\}^n$, the parity function χ_a is defined as $\chi_a(x) = (-1)^{\sum_{i=1}^n a_i x_i}$. It is known that any Boolean function $f : \{0,1\}^n \to \{-1,+1\}$ can be represented as a weighted sum of parity functions (see [7])

$$f(x) = \sum_{a \in \{0,1\}^n} \hat{f}(a)\chi_a(x)$$

where $\hat{f}(a) = \mathbf{E}_U[f(x)\chi_a(x)]$ is the *Fourier coefficient* of f at a. This is called the Fourier representation of f and is a direct consequence of the fact that $\{\chi_a \mid a \in \{0,1\}^n\}$ forms an orthonormal basis for all Boolean (or even real-valued) functions over $\{0,1\}^n$, i.e., $\mathbf{E}_U[\chi_a(x)\chi_b(x)]$ is one if $a = b$ and zero otherwise. Notice that if $f = \chi_c$ for some c then $\hat{f}(c) = 1$ and $\hat{f}(a) = 0$ for all $a \neq c$.

The focus of the paper is on a learning model in which the instance distribution is uniform and the noise process is characterized by a pair of parameters (D, R). The noise process can be viewed as drawing a random vector a from the distribution D (representing the attribute noise process) and a random value b from the distribution R (representing classification noise), then returning the exclusive OR of a with the original example vector x and the exclusive OR of the label $f(x)$ with b. So the noise process changes an example $(x, f(x))$ to an example $(x \oplus a, f(x) \oplus b)$ (actually, because we consider functions mapping to $\{-1,+1\}$, we will assume that R produces values in $\{-1,+1\}$ and replace the latter $\oplus$ with multiplication). We will call this (D, R)-noise and denote the oracle that returns a (D, R)-noisy example for f with respect to the uniform distribution by $EX_{D,R}(f, U)$.

Definition 1 *Let C be a concept class containing functions $f : \{0,1\}^n \to \{-1,+1\}$. Then C is* PAC learnable under the uniform distribution with (D, R)-noise *if there is an algorithm A such that for any $\epsilon, \delta \in (0,1)$ and for any target $f \in C$, given the inputs ϵ, δ and access to a noisy example oracle $EX_{D,R}(f, U)$, the algorithm A outputs a hypothesis h such that $\Pr_U[h \neq f] < \epsilon$ with probability at least $1 - \delta$. The algorithm must make a number of oracle calls (have* sample complexity*) at most polynomial in n, $1/\epsilon$, and $1/\delta$. If C can be learned for $\epsilon = 1/2 - 1/p(n)$, where $p(\cdot)$ is a fixed polynomial, then C is said to be* weakly learnable. *The* time complexity *of A is the number of computation steps taken by A. A PAC algorithm is* efficient *if its time complexity is also polynomial in n, $1/\epsilon$, and $1/\delta$.*

Notice that we are implicitly assuming in the definition above that D and R are known to the learning algorithm A, although as we will see later we can relax this somewhat for some of our positive results. However, in general, some assumption about the form of these noise distributions seems to be necessary. For example, consider the problem of learning the concept class M containing just the two monomials x_1 and $\overline{x_1}$ when the classification noise process R is arbitrary and no information about R is available to the learner. Then M is not learnable, because one cannot distinguish examples representing x_1 in a noiseless setting ($\mathbf{E}[R] = 1$) from examples representing $\overline{x_1}$ in a full-noise setting ($\mathbf{E}[R] = -1$). Thus, there are situations in which learning is possible if the noise process is known and impossible otherwise. If the classification noise process R always returns 1, then (D, R)-noise is simply attribute noise and we refer to it as D-noise. Our lower bounds focus on this type of noise.

3 Model Transformation

Before developing our main results, it is useful to relate the (D, R)-noise model to another model where the example $(x, f(x))$ is changed to $(x, f(x \oplus a)b)$ for a random vector a drawn according to distribution D and $b \in \{-1,+1\}$ drawn according to distribution R.

Lemma 1 *Let $U = U_n$ be the uniform distribution over n-bit vectors and (D, R) be any distributions over $\{0,1\}^n$ and $\{-1,+1\}$, respectively. Let $f : \{0,1\}^n \to \{-1,+1\}$ be any Boolean function. If $X \in_U \{0,1\}^n$, $A \in_D \{0,1\}^n$ and $B \in_R \{-1,+1\}$ are independent random variables then the random variables $(X \oplus A, f(X)B)$ and $(X, f(X \oplus A)B)$ have identical distributions.*

Proof Consider the random variables $X_1 = (X, A, B)$ and $X_2 = (X \oplus A, A, B)$. Since X is uniformly distributed and independent of A, X_1 and X_2 are identically distributed. Define $\varphi(x, y, z) = (x, f(x \oplus y)z)$. Then

$$(X \oplus A, f(X)B) = \varphi(X_2) \sim \varphi(X_1) = (X, f(X \oplus A)B),$$

where $\sim$ denotes that the two random variables have the same distribution. This completes the claim. □

This lemma is key to our subsequent results, as it allows us to consider the easier noise model of $(X, f(X \oplus A)B)$ instead of the random attribute noise model when learning is with respect to the uniform distribution.

4 Sample Complexity Lower Bound

In this section we give a lower bound for PAC-learning with D-noise. Because D-noise is a special case of (D, R)-noise, our lower bounds immediately generalize to this latter model as well.

We start with some intuition for the lower bound. Let C be the class being learned. Let f and g be two functions in the class C and suppose $\Pr_U[f \neq g] > \epsilon$. If for a fixed x and distribution D the expectation $\mathbf{E}_{a \sim D}[f(x \oplus a)]$ is very close to $\mathbf{E}_{a \sim D}[g(x \oplus a)]$, then we cannot notice the difference between $(x, f(x \oplus a_1))$ and $(x, g(x \oplus a_2))$. Now since the example oracle we consider chooses x according to the uniform distribution, we will look at $\mathbf{E}_x[|\mathbf{E}_a[f(x \oplus a) - g(x \oplus a)]|]$. This, we will show, is a good measure for learnability with noise. We now formalize the above.

Definition 2 *Let C be a concept class over $\{0,1\}^n$ and let $f, g \in C$. Let D be any distribution over $\{0,1\}^n$. Then the* noisy distance *between f and g under the distribution D is defined as*

$$\Delta_D(f, g) \equiv \frac{1}{2}\mathbf{E}_x[|\mathbf{E}_a[f(x \oplus a) - g(x \oplus a)]|],$$

where the expectation of x is taken over the uniform distribution over $\{0,1\}^n$ and the expectation of a is taken with respect to D. Also define for C and D as above and for any $\epsilon > 0$,

$$\Delta_D^\epsilon(C) \equiv \min\{\Delta_D(f, g) \mid f, g \in C \text{ with } \Pr_U[f \neq g] > \epsilon\}.$$

We say that f is ϵ-close to g (or vice versa) if $\Pr_U[f \neq g] \leq \epsilon$.

The following theorem states an information-theoretic lower bound on the number of samples required by any PAC learning algorithm.

Theorem 2 *Let C be a concept class and, for fixed ϵ and D, represent $\Delta_D^\epsilon(C)$ by Δ. Then any PAC learning algorithm for C under D-distribution noise that, with probability at least $1 - \delta/2$, outputs an $(\epsilon/2)$-close hypothesis requires a sample complexity of $\Omega\left(\frac{1-\delta}{\Delta}\right)$.*

Proof Consider an algorithm that tries to distinguish whether a sample $S = \{\langle x_i, b_i\rangle \mid i \in [m]\}$ is labeled by the function f or g, where $f, g \in C$ and $\Delta_D(f,g) = \Delta$. The claim is that no algorithm has a distinguishing probability greater than $2m\Delta$.

Formally, let F and G be distributions over $\{0,1\}^n \times \{-1,+1\}$ that produce $\langle x, f(x \oplus a)\rangle$ and $\langle x, g(x \oplus a)\rangle$, respectively, where x is drawn according to the uniform distribution and a is drawn according to the noise distribution D. Also let F^m and G^m be induced distributions on m independent samples drawn according to F and G, respectively. We will show that there exists no (possibly randomized) prediction algorithm A (that outputs $\{0,1\}$) with the property that

$$\left| \Pr_{S \sim F^m, r}[A(S) = 1] - \Pr_{S \sim G^m, r}[A(S) = 1] \right| > 2m\Delta,$$

where r denotes the randomness of A.

This relates to the PAC confidence parameter δ as follows. Fix any prediction algorithm A and denote by $\delta_A^m(f,g)$ the above absolute difference of probabilities. Then the probability that A correctly predicts whether the sample was drawn from F^m or G^m is at most $\delta_A^m(f,g) + (1 - \delta_A^m(f,g))/2$. That is, with probability $\delta_A^m(f,g)$, A can distinguish the source of the sample, and in the best case it predicts correctly every time this occurs. However, with probability $1 - \delta_A^m(f,g)$, A can at best guess randomly as to whether the source is F^m or G^m. Therefore, since below we will show that, for all A, $\delta_A^m(f,g) \leq 2m\Delta$, we have that if any algorithm A correctly predicts whether the sample source is F^m or G^m with probability at least $1/2 + \delta'/2$ then the sample size m used must be such that $m \geq \delta'/(2\Delta)$. Making the substitution $\delta' = 1 - \delta$ gives the PAC form of the bound stated in the theorem. So what remains is to prove the the bound on $\delta_A^m(f,g)$.

Let $F(x,y), G(x,y)$ be the probability weight assigned to $(x,y) \in \{0,1\}^n \times \{-1,+1\}$ by F, G, respectively. Note that $F(x,y) = \frac{1}{2^{n+1}}(1 + \mathbf{E}_a[y \cdot f(x \oplus a)])$, which implies that for all x, y,

$$|\mathbf{E}_a[f(x \oplus a) - g(x \oplus a)]| = 2^{n+1}|F(x,y) - G(x,y)|.$$

Relating this to the noisy distance between f and g, we have $\Delta_D(f,g) = \sum_x |F(x,1) - G(x,1)|$ (notice that for every x, $|F(x,y) - G(x,y)|$ is independent of the value of y).

Now we define $\Delta(F,G) = \sum_{x,y} |F(x,y) - G(x,y)|$, so we have that $\Delta(F,G) = 2\Delta_D(f,g) = 2\Delta$. Notice that $\Delta(F,G)$ is a measure of the distance between two probability distributions F and G in terms of the L_1 norm of the difference of these distributions viewed as vectors over $\{0,1\}^n \times \{-1,+1\}$. More generally, for $m \geq 1$ define $\Delta(F^m, G^m) = \sum_{\vec{x},\vec{y}} |F^m(\vec{x},\vec{y}) - G^m(\vec{x},\vec{y})|$, where $\vec{x} \in (\{0,1\}^n)^m$ and $\vec{y} \in \{-1,+1\}^m$. This measure is similar to the *statistical distance* of Yang [10], although his measure uses an L_2 norm. We will now use an approach similar to Yang's, based on the subadditivity of our distance measure, to obtain our result.

First, notice that for all $m \geq 1$, $\Delta(F^m, G^m)$ is an upper bound on $\delta_A^m(f,g)$, since

$$\begin{aligned}
\left| \Pr_{S \sim F^m, r}[A(S) = 1] - \Pr_{S \sim G^m, r}[A(S) = 1] \right| &= \left| \sum_{\vec{x},\vec{y}} \Pr_r[A(\vec{x},\vec{y}) = 1] \cdot (F^m(\vec{x},\vec{y}) - G^m(\vec{x},\vec{y})) \right| \\
&\leq \sum_{\vec{x},\vec{y}} |F^m(\vec{x},\vec{y}) - G^m(\vec{x},\vec{y})|
\end{aligned}$$

We prove next that $\Delta(\cdot,\cdot)$ is subadditive, that is, that $\Delta(F^m, G^m) \leq \Delta(F^{m-1}, G^{m-1}) + \Delta(F,G)$.

Let $x \in \{0,1\}^n$, $y \in \{-1,+1\}$, $\alpha \in (\{0,1\}^n)^{m-1}$, and $\beta \in \{-1,+1\}^{m-1}$, for $m > 1$. Then

$$\begin{aligned}
\Delta(F^m, G^m) &= \sum_{x,y,\alpha,\beta} |F(x,y)F^{m-1}(\alpha,\beta) - G(x,y)G^{m-1}(\alpha,\beta)| \\
&\leq \sum_{x,y,\alpha,\beta} |F(x,y)F^{m-1}(\alpha,\beta) - F(x,y)G^{m-1}(\alpha,\beta)| \\
&\quad + \sum_{x,y,\alpha,\beta} |F(x,y)G^{m-1}(\alpha,\beta) - G(x,y)G^{m-1}(\alpha,\beta)| \\
&\leq \sum_{\alpha,\beta} |F^{m-1}(\alpha,\beta) - G^{m-1}(\alpha,\beta)| + \sum_{x,y} |F(x,y) - G(x,y)|.
\end{aligned}$$

Thus, $\delta_A^m(f,g) \leq \Delta(F^m, G^m) \leq m\Delta(F,G) = 2m\Delta$, proving the theorem. □

4.1 Near Tight Characterization

In the following we will use Fourier analysis to obtain a nearly tight characterization of the noisy distance quantity $\Delta_D(f,g)$.

Definition 3 *Let $f : \{0,1\}^n \to \{-1,+1\}$ be a Boolean function and let $\alpha \in [0,1]^{\{0,1\}^n}$ be a vector of reals in the range $[0,1]$ indexed by n-bit vectors. Then the* α-attenuated power spectrum *of f is*

$$s_\alpha(f) = \sum_c \alpha_c^2 \hat{f}(c)^2.$$

As it turns out, $\Delta_D(f,g)$ is characterized by the α-attenuated power spectrum of $f - g$ when α is defined as follows:

$$\alpha_c = \mathbf{E}_{a \sim D}[\chi_c(a)]. \tag{1}$$

In particular, define $s_D(f)$ to be $s_\alpha(f)$ with α_c defined in this way. Then we have:

Theorem 3 *Let $f, g : \{0,1\}^n \to \{-1,+1\}$ be Boolean functions and D any probability distribution over $\{0,1\}^n$. Then*

$$\frac{s_D(f-g)}{4} \leq \Delta_D(f,g) \leq \sqrt{s_D(f-g)}. \tag{2}$$

Proof Using the fact that $\mathbf{E}[|X|] \leq \sqrt{\mathbf{E}[X^2]}$, we get

$$\Delta_D(f,g) \leq \frac{1}{2}\sqrt{\mathbf{E}_{x \sim U_n}[(\mathbf{E}_{a \sim D}[f(x \oplus a) - g(x \oplus a)])^2]}.$$

Let $h(x) = (f(x) - g(x))/2$. Then right hand side of the previous expression becomes

$$\sqrt{\mathbf{E}_x[\mathbf{E}_a^2[h(x \oplus a)]]}.$$

We now work with the inner expression $\mathbf{E}_x[\mathbf{E}_a^2[h(x \oplus a)]]$.

$$\begin{aligned}
\mathbf{E}_x[\mathbf{E}_a^2[h(x \oplus a)]] &= \mathbf{E}_x[\mathbf{E}_a[h(x \oplus a)]\mathbf{E}_b[h(x \oplus b)]] \\
&= \mathbf{E}_{a,b}\left[\mathbf{E}_x\left[\sum_{s,t} \hat{h}(s)\hat{h}(t)\chi_s(x \oplus a)\chi_t(x \oplus b)\right]\right] \\
&= \mathbf{E}_{a,b}\left[\sum_{s,t} \hat{h}(s)\hat{h}(t)\chi_s(a)\chi_t(b)\mathbf{E}_x[\chi_s(x)\chi_t(x)]\right] \\
&= \sum_s \hat{h}(s)^2\mathbf{E}_a^2[\chi_s(a)] \\
&= s_D(h).
\end{aligned}$$

Hence we get

$$\Delta_D(f,g) \leq \sqrt{s_D(f-g)}.$$

Next, we show a lower bound on $\Delta_D(f,g)$. We note that $0 \leq |\mathbf{E}_a[h(x \oplus a)]| \leq 1$, since $h \in \{-1,0,+1\}$. Thus

$$\Delta_D(f,g) = \mathbf{E}_x[|\mathbf{E}_a[h(x \oplus a)]|] \geq \mathbf{E}_x[\mathbf{E}_a^2[h(x \oplus a)]] = s_D(h) = \frac{s_D(f-g)}{4}$$

using the same analysis as in the upper bound. This completes the theorem. □

Define

$$S_D^\epsilon(C) = \min\{s_D(f-g) \mid f,g \in C \text{ with } \Pr_U[f \neq g] > \epsilon\}.$$

Using this definition with Theorem 3 we have the following inequalities.

Theorem 4 *For any class C and any ϵ we have*

$$\frac{S_D^\epsilon(C)}{4} \leq \Delta_D^\epsilon(C) \leq \sqrt{S_D^\epsilon(C)}.$$

Then combining this with Theorem 2 we have the following lower bound.

Theorem 5 *Let C be a concept class with $S_D^\epsilon(C) \leq S$. Then any PAC learning algorithm for C under D-distribution attribute noise that outputs an $(\epsilon/2)$-good hypothesis with probability at least $1-\delta/2$ requires a sample complexity of $\Omega\left(\frac{1-\delta}{\sqrt{S}}\right)$.*

We now show that the class of parity functions is not PAC learnable under the uniform distribution with D-noise for almost every noise distribution D.

Theorem 6 *Let D be a distribution such that $\max_a D(a)$ is superpolynomially small (or $1/\omega(poly(n))$). Then the set of parity functions is not PAC-learnable under D-distribution noise.*

Proof Notice that for any two distinct parity functions f and g we have $\Pr[f \neq g] = 1/2$. Since f and g are parity functions, $s_D(f-g) = s_D(f) + s_D(g)$, and by the preceding theorem it is enough to find two distinct parity functions f and g with superpolynomially small $s_D(f)$ and $s_D(g)$.

First, notice that

$$\alpha_c = \mathbf{E}_{a\sim D}[\chi_c(a)] = \sum_a \chi_c(a)D(a) = 2^n\hat{D}(c).$$

Also, by Parseval's identity,

$$\sum_c \hat{D}^2(c) = \mathbf{E}_a[D^2(a)].$$

Therefore,

$$\mathbf{E}_{c\sim U_n}[s_D(\chi_c)] = \mathbf{E}_c[\alpha_c^2] = \sum_a D^2(a) \leq [\max_a D(a)]\sum_a D(a) \leq \max_a D(a).$$

Thus, since $s_D(f)$ is nonnegative for all D and Boolean f, only a superpolynomially small fraction of parity functions χ_c can have $s_D(\chi_c)$ inverse polynomially large if $D(x)$ is superpolynomially small for all x. So there are at least two parity functions f and g for which both $s_D(f)$ and $s_D(g)$ are superpolynomially small. □

Finally, it should be noted that Theorem 5 is only a hardness result for strong PAC learnability. As an example of a class that can be weakly learned in spite of arbitrary and unknown random attribute noise, consider monotone Boolean functions. Blum, Burch, and Langford [1] have shown that every monotone Boolean function f is weakly approximated with respect to the uniform distribution by either one of the two constant functions or by the majority function. Since applying random attribute noise alone does not change the expected value of the label of f, f is weakly approximated by a constant function if and only if the noisy function represented by $EX_D(f,U)$ is weakly approximated by a constant function. This implies an obvious algorithm for weakly learning monotone functions with respect to the uniform distribution despite arbitrary unknown attribute noise.

5 Upper Bounds

In this section we consider a certain type of Fourier-based learning algorithm which we will call *LMN-style*. The LMN-style algorithm was introduced by Linial, Mansour, and Nisan [7], who showed that the class AC^0 of polynomial-size, constant depth circuits is PAC learnable with respect to the uniform distribution in quasipolynomial (roughly $n^{\text{polylog}(n)}$) time. The key to their result was analyzing the Fourier properties of AC^0 to show that for every AC^0 function f, the sum of the squares of the Fourier coefficients of degree polylog(n) or less is nearly 1. They then showed that the function

$$h(x) = \text{sign}\left(\sum_{|a|\leq \text{polylog}(n)} \hat{f}(a)\chi_a(x)\right)$$

is a good approximator to the target function f (here $|a|$ denotes the Hamming weight of a, that is, the number of 1's it contains). Finally, it follows from standard Hoeffding bounds that all of these Fourier coefficients can be closely approximated by sampling from a uniform-distribution example oracle, with sample size and running time dominated by $n^{\text{polylog}(n)}$.

An LMN-style algorithm, then, given $\epsilon > 0$, consists of estimating—for every n-bit index in a set T_ϵ—Fourier coefficients, with the guarantee that the sum of the squares of these coefficients is

at least $1-\epsilon$. For example, in the case of Linial *et al.*'s algorithm for AC^0, the Hamming weight of the Fourier indices in T_ϵ grows as ϵ approaches 0. The hypothesis resulting from an LMN-style algorithm will be of the form

$$h(x) = \text{sign}\left(\sum_{a \in T_\epsilon} \tilde{f}(a)\chi_a(x)\right),$$

where $\tilde{f}(a)$ represents an estimate of the Fourier coefficient $\hat{f}(a)$.

In this section we show that if there is an LMN-style algorithm for learning a class of functions C, then C is PAC-learnable under any (D,R)-noise in time polynomial in $|T|$, $1/(1-2\eta)$, and $1/\Delta_D^\epsilon(AC^0)$, where η is the expectation of the noise rate in the label (*i.e.*, $\eta = \mathbf{E}[(1-R)/2]$). Since $1/\Delta_D^\epsilon(AC^0)$ is a lower bound for PAC-learning with D-distribution noise and $1/(1-2\eta)$ is a lower bound for learning with label noise [9], our result is tight (up to polynomial factors). Before we formally state the result, we recall the following version of Hoeffding bounds.

Lemma 7 *(Hoeffding bounds) Let X_i, $1 \le i \le m$, be independent, identically distributed random variables, where $\mathbf{E}[X_i] = \mu$ and $|X_i| \le B$. Then*

$$\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| > \gamma\right] \le \delta,$$

whenever $m \ge (2B^2/\gamma^2)\ln(2/\delta)$.

Theorem 8 *Let C be a class of Boolean functions that is closed under complement and suppose that C is learnable with respect to the uniform distribution by an LMN-style algorithm using index set T_ϵ. Then for every $\epsilon > 0$ for which the set of parity functions indexed by T_ϵ is a subset of C, C is learnable with respect to the uniform distribution under any known (D,R)-noise in time polynomial in $1/\epsilon, 1/\delta, 1/(1-2\eta), |T_\epsilon|$, and $1/\Delta_D^\epsilon(C)$, where η is the expectation of the classification noise rate.*

Proof Let $\Delta = \Delta_D^\epsilon(C)$ and $T = T_\epsilon$. Recall from (1) that in defining $s_D(f)$, $\alpha_c = \mathbf{E}_{a\sim D}[\chi_c(a)]$. First we note that there is at most one $c \in T$ such that $|\alpha_c| < \Delta/2$, since otherwise there are distinct c_1 and c_2 such that $|\alpha_{c_1}| < \Delta/2$, $|\alpha_{c_2}| < \Delta/2$. This implies by Theorem 3 that

$$\Delta^2 \le s_D(\chi_{c_1} - \chi_{c_2}) = s_D(\chi_{c_1}) + s_D(\chi_{c_2}) = \alpha_{c_1}^2 + \alpha_{c_2}^2 \le \Delta^2/2,$$

which is a contradiction. So let $c_0 \in T$, if it exists, be the unique index such that $|\alpha_{c_0}| < \Delta/2$. Actually, we will now argue that there is no such c_0 in T. Since C is closed under complement, if $c_0 \in T$, then $-\chi_{c_0} \in C$. Thus $|\mathbf{E}_{a\sim D}[\chi_{c_0}(a)]| = |\mathbf{E}_{a\sim D}[-\chi_{c_0}(a)]| < \Delta/2$, which contradicts the uniqueness of c_0.

The rest of the proof applies the standard LMN analysis [7] after adjusting against the effects of the error rates. To find the Fourier coefficient $\hat{f}(c)$ of f at $c \in T$, we take a sample S_m of size m (to be determined later), $S_m = \{(x^i \oplus a^i, f(x^i)b^i) \mid 1 \le i \le m\}$ (since f is $\{\pm 1\}$-valued, we choose $b \in \{-1,+1\}$, so XOR becomes multiplication), and estimate the expectation $\mu_c = \mathbf{E}_{x,a,b}[f(x)b\chi_c(x \oplus a)]$. Note that

$$\begin{aligned}
\mu_c &= \mathbf{E}_{x\sim U_n}[\mathbf{E}_{a\sim D}[\mathbf{E}_{b\sim U}[f(x)b\chi_c(x\oplus a)]]] \\
&= \mathbf{E}_b[b]\mathbf{E}_x[f(x)\chi_c(x)]\mathbf{E}_a[\chi_c(a)] \\
&= (1-2\eta)\hat{f}(c)\alpha_c.
\end{aligned}$$

Because we are assuming that D and R are known, the factors of $(1-2\eta)$ and α_c are known and can easily be eliminated. We assume that η and the α_c's are exactly known; a more tedious error analysis could be done to eliminate this assumption and is given in Appendix A.

Thus, for each $c \in T$, a good estimate of μ_c gives a good estimate of the Fourier coefficient $\hat{f}(c)$. So, for a fixed $c \in T$, let $\beta_c = \frac{1}{m}\sum_{i=1}^{m} \chi_c(x^i \oplus a^i) f(x^i) b^i$ be the estimate for μ_c. Using the Hoeffding bound of Lemma 7, we can estimate this expectation with a sample size (and time complexity, with polynomial blowup) of

$$m = \frac{32|T|}{\epsilon(1-2\eta)^2\Delta^2} \ln \frac{4|T|}{\delta}$$

(i.e., letting $B = 1$, $\gamma = \sqrt{\epsilon/(2|T|)}(1-2\eta)\Delta/2$, and using $\delta/(2|T|)$ as the confidence). This will guarantee that with probability at least $1 - \delta/2$, $|\beta_c - \mu_c| < \sqrt{\epsilon/(2|T|)}(1-2\eta)|\alpha_c|$ holds simultaneously for all $c \in T$. This in turn implies with the same probability that for all c, $|\hat{\beta}_c - \hat{f}(c)| < \sqrt{\epsilon/(2|T|)}$, where $\hat{\beta}_c = \beta_c(1-2\eta)^{-1}\alpha_c^{-1}$ is the estimate for $\hat{f}(c)$. This shows that the set L of all of the relevant coefficients indexed by T can be estimated in time polynomial in $|T|$, $1/\Delta$, $1/(1-2\eta)$, $1/\epsilon$, and $1/\delta$. Letting $g(x) = \sum_{c \in T} \hat{\beta}_c \chi_c(x)$, the final hypothesis is $h(x) = \mathrm{sign}(g(x))$. By the standard LMN analysis, we get

$$\begin{aligned}
\Pr_x[\mathrm{sign}(g(x)) \neq f(x)] &\leq \frac{1}{4}\mathbf{E}_x[(f(x)-g(x))^2] = \frac{1}{4}\sum_c (\hat{f}(c) - \hat{g}(c))^2 \\
&= \frac{1}{4}\left[\sum_{c \notin T_\epsilon} \hat{f}(c)^2 + \sum_{c \in T_\epsilon} (\hat{f}(c) - \hat{\beta}_c)^2\right] \\
&< \epsilon.
\end{aligned}$$

□

For the LMN-style algorithm for AC^0, as long as $1/\epsilon = O(n^{\mathrm{polylog}(n)})$, the parity functions indexed by T_ϵ are of polylogarithmic degree (by results in Linial *et al.* [7]) and are therefore in AC^0 since parity on polylogarithmic bits can be computed in AC^0 by a result of Håstad (see [5], where Theorem 2.2, page 13, proved that Parity can be computed by circuits of size $O(n^{\frac{d-2}{d-1}} 2^{n^{1/(d-1)}})$ and depth d; so, a Parity on $O((\log_2 n)^{d-1})$ inputs can be computed in $n^{O(1)}$ size and depth d). This immediately gives us the following result.

Theorem 9 *For $1/(1-2\eta) = O(n^{\mathrm{polylog}(n)})$, $1/\epsilon = O(n^{\mathrm{polylog}(n)})$, and $1/\delta = O(2^n)$, the class AC^0 of constant depth, polynomial size circuits is learnable under the uniform distribution with any known (D,R)-noise in time dominated by*

$$n^{\mathrm{polylog}(n)} p(1/\Delta_D^\epsilon(AC^0))$$

where $p(\cdot)$ is a fixed polynomial.

As a specific example of the application of this theorem, consider a known attribute noise process D that is a product distribution over $\{0,1\}^n$ defined by possibly distinct noise rates $0 \leq p_i \leq 1$ for each attribute $1 \leq i \leq n$. That is, a vector a is chosen according to D by independently setting each element a_i to 1 with probability p_i. We claim that if $p_i = O(1/\mathrm{polylog}(n))$ for all i (and if other parameters are bounded as in Theorem 9) then there is a learning algorithm for AC^0 with

time dominated by $n^{\text{polylog}(n)}$. To see this, recall that the hypothesis in an LMN-style algorithm is formed using only (estimates of) coefficients indexed by T_ϵ, and that for AC^0 all of these indices have polylogarithmic (in n) Hamming weight when ϵ is as given in Theorem 9. Furthermore, based on results of Linial *et al.* [7] (see the analysis at the end of the proof of Theorem 8), if f and g are AC^0 functions such that

$$\Pr[f \neq g] > \epsilon = \Omega(1/n^{\text{polylog}(n)})$$

then the difference $\hat{f}(c) - \hat{g}(c)$ must be at least $1/n^{\text{polylog}(n)}$ large for at least one of the coefficients indexed by T_ϵ. But then $s_D(f-g) = \sum_c \alpha_c^2 (\widehat{f-g})^2(c) = \sum_c \alpha_c^2 (\hat{f}(c) - \hat{g}(c))^2$ (the final equality follows by linearity of the Fourier transform) will be inverse quasipolynomially large as along as $\alpha_c = \mathbf{E}_{a \sim D}[\chi_c(a)]$ is inverse quasipolynomial for all c in T_ϵ. A simple probabilistic analysis shows that in fact all of these α_c will be sufficiently large as long as $|c|$ is polylogarithmic in n. In particular,

$$\begin{aligned}
\alpha_c &= \mathbf{E}_{a \sim D}\left[(-1)^{\sum_{i=1}^n a_i c_i}\right] \\
&= \prod_{i=1}^n \mathbf{E}_{a_i \sim D_i}\left[(-1)^{a_i c_i}\right], \text{ since } D \text{ is a product distribution} \\
&= \prod_{i \in c} (1 - 2p_i) \\
&> (1 - 1/\text{polylog}(n))^{|c|}, \text{ since } (\forall i)\ p_i < 1/\text{polylog}(n) \\
&> 1/n^{\text{polylog}(n)}, \quad \text{since } |c| \leq \text{polylog}(n)
\end{aligned}$$

Therefore, for attribute noise D as defined, $\Delta_D^\epsilon(AC^0)$ is inverse quasipolynomially large, and our claim follows by the preceding theorem.

6 Conclusion

In this work, we have studied noisy learning models under the uniform distribution. We showed that Fourier analysis is useful in this setting even in cases where both attribute and classication noise are present. The Fourier analysis used in this work led to a natural parameter that was used to characterize the upper and lower bounds for learning complexity in the noisy model.

It would be interesting to explore the extent to which our techniques can be generalized to non-uniform distributions. In addition, we showed that under certain conditions, AC^0 is learnable despite attribute noise; are there other natural concept classes where our techniques can be used to show that learning remains possible despite attribute noise? Finally, the existing connections between our techniques and the work done in [2] merit further investigation.

Acknowledgments

We thank the anonymous referees for their generous and constructive comments on this paper. In fact, the comments of one particular referee led to significant simplifications of two of our original proofs.

The second author acknowledges that this material is based upon work supported by the National Science Foundation under Grants No. CCR-9800029, CCR-9877079, and CCR-0209064.

References

[1] Blum, A., Burch, C., and Langford, J. (1998), On Learning Monotone Boolean Functions, *in* Proc. 39th Ann. IEEE Symp. Foundations of Computer Science, pp. 408–415.

[2] Benjamini, I., Kalai, G., and Schramm, O. (1999), Noise Sensitivity of Boolean Functions and Applications to Percolation, *Inst. Hautes Études Sci. Publ. Math.*, **90**, pp. 5-43.

[3] Decatur, S., and Gennaro, R. (1995), On Learning from Noisy and Incomplete Examples, *in* Proc. 8th Ann. ACM Conf. Computational Learning Theory, pp. 353–360.

[4] Goldman, S., and Sloan, S (1995), Can PAC Learning Algorithms Tolerate Random Attribute Noise?, *Algorithmica*, **14**, 1, pp. 70-84.

[5] Håstad, J. (1987), *Computational Limitations for Small-Depth Circuits*, The MIT Press.

[6] Kahn, J., Kalai, G., and Linial, N. (1988), The Influence of Variables on Boolean Functions, *in* Proc. 29th Ann. Symp. Foundations of Computer Science, pp. 68-80.

[7] Linial, N., Mansour, Y., and Nisan, N. (1993), Constant Depth Circuits, AC^0 Circuits, and Learnability, *Journal of the ACM*, **40**, 3, pp. 607-620.

[8] Shackelford, G., and Volper, D. (1988), Learning k-DNF with Noise in the Attributes, *in* Proc. 1988 Workshop on Computational Learning Theory, pp. 97-103.

[9] Simon, H.U. (1993), General Bounds on the Number of Examples Needed for Learning Probabilistic Concepts, *in* Proc. 6th Ann. ACM Workshop on Computational Learning Theory, pp. 402-411.

[10] Yang, K. (2001), On Learning Correlated Boolean Functions using Statistical Queries, *in* Proc. 12th Int. Conf. Algorithm Learning Theory, pp. 59-76.

A Error Analysis

We make the standard assumption that $\eta \leq \eta_b < 1/2$ for some fixed known bound η_b. Let

$$\kappa = \frac{1}{32}\sqrt{\frac{\epsilon}{|T|}}, \quad M = \min\left\{1 - 2\eta_b, \frac{\Delta}{2}\right\}.$$

Next we set $\varrho = \kappa \cdot M$. Note that clearly $\varrho < \Delta/4$.

Assume that we have approximations $\hat{\eta}$ to the noise rate η and $\hat{\alpha}_c$ to α_c, for all relevant c's. More specifically, suppose we have

$$|\eta - \hat{\eta}| < \varrho \quad \text{and} \quad |\alpha_c - \hat{\alpha}_c| < \varrho, \ \forall c \in T$$

with probability at least $1 - \delta/2$. These approximations might be obtained by sampling from some sort of oracles for D and R, for example. Note that

$$\frac{\varrho}{1 - 2\eta} \leq \kappa, \quad \text{and} \quad \frac{\varrho}{|\alpha_c|} \leq \kappa$$

Consider the m-sample $S_m = \{(\tilde{x}^i, \tilde{b}^i) \mid i \in [m]\}$, where $\tilde{x}$ is the a-noisy attribute vector and $\tilde{b}$ is the η-noisy label. From S_m, we estimate the quantity

$$\beta_c = \frac{1}{m} \sum_{i=1}^{m} \chi_c(\tilde{x}^i)\tilde{b}^i$$

whose expectation is $\mu_c = (1-2\eta)\alpha_c \hat{f}(c)$. Since our goal is to estimate $\hat{f}(c)$ without knowing precisely the values of η and α_c's, we consider the estimate

$$\tilde{\beta}_c = \frac{\beta_c}{(1-2\hat{\eta})\hat{\alpha}_c}$$

The expectation of this estimate is

$$\mathbf{E}[\tilde{\beta}_c] = \frac{1-2\eta}{1-2\hat{\eta}} \frac{\alpha_c}{\hat{\alpha}_c} \hat{f}(c) \tag{3}$$

Claim 10 *For some $\vartheta < 16\kappa$,*

$$\mathbf{E}[\tilde{\beta}_c] = (1 \pm \vartheta)\hat{f}(c)$$

Proof First we note

$$|\eta - \hat{\eta}| < \varrho \Longrightarrow |(1-2\eta) - (1-2\hat{\eta})| < 2\varrho \Longrightarrow \frac{1-2\eta}{1-2\eta+2\varrho} < \frac{1-2\eta}{1-2\hat{\eta}} < \frac{1-2\eta}{1-2\eta-2\varrho}$$

and thus

$$1 - \frac{\xi}{1+\xi} \;<\; \frac{1-2\eta}{1-2\hat{\eta}} \;<\; 1 + \frac{\xi}{1-\xi}, \quad \text{where} \quad \xi = \frac{2\varrho}{1-2\eta} \leq 2\kappa.$$

Applying a similar manipulation to α_c we get

$$1 - \frac{\tau}{1+\tau} \;<\; \frac{\alpha_c}{\hat{\alpha}_c} \;<\; 1 + \frac{\tau}{1-\tau}, \quad \text{where} \quad \tau = \frac{2\varrho}{\alpha_c} \leq 2\kappa.$$

Thus Equation 3 becomes

$$\left(1 - \frac{\xi}{1+\xi}\right)\left(1 - \frac{\tau}{1+\tau}\right)\hat{f}(c) \;<\; \mathbf{E}[\tilde{\beta}_c] \;<\; \left(1 + \frac{\xi}{1-\xi}\right)\left(1 + \frac{\tau}{1-\tau}\right)\hat{f}(c)$$

This yields (after dropping and adding some terms)

$$\left(1 - \left[\frac{\xi}{1+\xi} + \frac{\tau}{1+\tau}\right]\right)\hat{f}(c) \;<\; \mathbf{E}[\tilde{\beta}_c] \;<\; \left(1 + 2\left[\frac{\xi}{1-\xi} + \frac{\tau}{1-\tau}\right]\right)\hat{f}(c)$$

We can simplify this further by setting $\theta = \xi/(1-\xi) + \tau/(1-\tau)$ and getting

$$(1-2\theta)\hat{f}(c) \;<\; \mathbf{E}[\tilde{\beta}_c] \;<\; (1+2\theta)\hat{f}(c) \;\Longrightarrow\; |\mathbf{E}[\tilde{\beta}_c] - \hat{f}(c)| \;<\; 2\theta\hat{f}(c).$$

Finally set $\vartheta = 2\theta$.

Note that $x/(1-x) \leq 2x$, for $x \in (0, 1/2)$, and hence $\vartheta \leq 4(\xi + \tau) \leq 16\kappa$. □

Now using Hoeffding bounds, to guarantee that

$$\Pr\left[\left|\tilde{\beta}_c - \mathbf{E}[\tilde{\beta}_c]\right| \geq \sqrt{\frac{\epsilon}{4|T|}}\right] \;\leq\; \frac{\delta}{2|T|}$$

we need to take

$$m \;=\; \frac{128|T|}{\epsilon(1-2\hat{\eta})^2\Delta^2}\;\ln\frac{4|T|}{\delta}$$

since $|\alpha_c| \geq \Delta/2$ and $|\alpha_c - \hat{\alpha}_c| < \varrho < \Delta/4$.

This implies that with probability at least $1-\delta/2$, we have $|\tilde{\beta}_c - (1 \pm \vartheta)\hat{f}(c)| < \sqrt{\epsilon/(4|T|)}$, for all $c \in T$. Thus

$$\sum_{c\in T}(\tilde{\beta}_c - \hat{f}(c))^2 \leq \sum_{c\in T}\left(\vartheta\hat{f}(c) + \sqrt{\frac{\epsilon}{4|T|}}\right)^2 \leq \epsilon$$

since $|\vartheta| \leq \sqrt{\epsilon/(4|T|)}$ and $|\hat{f}(c)| \leq 1$.