Uniform-Distribution Learnability of Noisy Linear Threshold Functions with Restricted Focus of Attention

Jeffrey C. Jackson*

Mathematics and Computer Science Dept.
Duquesne University, Pittsburgh PA 15282-1754 USA
jacksonj@duq.edu

Abstract. Recently, Kalai *et al.* [1] have shown (among other things) that linear threshold functions over the Boolean cube and unit sphere are agnostically learnable with respect to the uniform distribution using the hypothesis class of polynomial threshold functions. Their primary algorithm computes monomials of large constant degree, although they also analyze a low-degree algorithm for learning origin-centered halfspaces over the unit sphere. This paper explores noise-tolerant learnability of linear thresholds over the cube when the learner sees a very limited portion of each instance. Uniform-distribution weak learnability results are derived for the agnostic, unknown attribute noise, and malicious noise models. The noise rates that can be tolerated vary: the rate is essentially optimal for attribute noise, constant (roughly 1/8) for agnostic learning, and non-trivial ($\Omega(1/\sqrt{n})$) for malicious noise. In addition, a new model that lies between the product attribute and malicious noise models is introduced, and in this stronger model results similar to those for the standard attribute noise model are obtained for learning homogeneous linear thresholds with respect to the uniform distribution over the cube. The learning algorithms presented are simple and have small-polynomial running times.

1 Introduction

A linear threshold function over the Boolean cube $\{0,1\}^n$ is any function that can be defined by taking the sign of the sum of a constant threshold value plus the dot product of a fixed vector of weights and the vector of the function's inputs. While the class of linear threshold functions can be learned in polynomial time with respect to arbitrary distributions over the cube (by using any polynomial-time linear programming solver), many open questions remain concerning the learnability of linear thresholds in the presence of noise.

Significant progress on noise-tolerant learning of linear thresholds was made recently when Kalai *et al.* [1] showed (among other things) that linear threshold functions over the Boolean cube are agnostically learnable with respect to

* This material is based upon work supported by the National Science Foundation under Grant No. CCR-0209064.

G. Lugosi and H.U. Simon (Eds.): COLT 2006, LNAI 4005, pp. 304–318, 2006.

the uniform distribution $\mathcal{U}_n$. Specifically, their algorithm, given $\epsilon > 0$ and a uniform-distribution oracle for any function $f : \{0,1\}^n \rightarrow \{-1,1\}$, produces an approximator h such that $\Pr_{\mathcal{U}_n}[f \neq h]$ is at most ϵ greater than the minimal error of any linear threshold used as an approximator to f. As there are relatively few positive results for the agnostic learning model, it is perhaps somewhat surprising that a positive result could be obtained for such a rich class.

Kalai *et al.*'s *polynomial regression algorithm*, while polynomial-time for constant ϵ, produces as its hypothesis h a large-constant-degree polynomial threshold function. Furthermore, to produce this hypothesis, the algorithm uses estimates of Fourier coefficients of the target f that involve computing monomials of degree up to d over the examples, where d is a large constant. While Kalai *et al.* also show that a degree-1 version of their algorithm produces reasonably good agnostic results when learning over the unit sphere, there is not an obvious translation of their analysis to uniform-distribution learning over the discrete cube.

This paper considers uniform-distribution learning of noisy linear thresholds over the Boolean cube when the learner is restricted to look at only a very few bits k of each example. This Restricted Focus of Attention (k-RFA) model was introduced by Ben-David and Dichterman [2] and has been considered in several settings. One reason for considering this model is that, when positive RFA results are possible, the resulting learning algorithms may be—and are, in this paper—relatively simple and efficient, since they are using relatively little information in each example.

In addition, there are theoretical reasons to be particularly interested in RFA learnability of linear thresholds. It has long been known that the Chow parameters of a linear threshold function f over the cube—parameters which can be efficiently estimated while looking at only one input bit plus the label per example—provide a unique signature for f: no other Boolean function has exactly the same Chow parameters. Thus, noiseless linear thresholds are information-theoretically learnable in the 1-RFA model. It is therefore natural to ask how much we can learn about a *noisy* linear threshold function given a similarly limited amount of information.

Algorithms are presented for RFA-learning linear threshold functions over the cube with respect to the uniform distribution in several noise models (described later): a weak version of agnostic learning, attribute noise generated by an unknown noise process, malicious noise, and a new model called restricted context-sensitive attribute noise (RCSAN, pronounced arc-san). In this model, unlike attribute noise, the noise process is allowed to specify multiple noise rates for an attribute, with the choice of rate for an example $(x, f(x))$ based on the values of a restricted set of attributes of x as well as the label $f(x)$. This generalizes the product version of the attribute noise model, in which noise is applied to each attribute i of an example independently at rate p_i.

In each of these models, our algorithms produce approximating hypotheses that (with high probability) agree with the target function with probability at least $1/2+\gamma$ for some γ that depends on how far the actual noise rate falls below the limits given next. For agnostic learning, any function f for which the optimal linear threshold function has error non-negligibly less than 1/8 can be efficiently weakly learned. For unknown attribute noise applied to a linear threshold function, weak learning is achieved for any noise process for which the marginal attribute noise rates are all non-negligibly less than 1/2. For malicious noise, a rate of $\Omega(1/\sqrt{n})$ can be tolerated. We consider several constrained versions of the RCSAN model. Our strongest RCSAN result shows, roughly speaking, that homogeneous linear threshold functions (linear thresholds that have a constant threshold value of 0) can be 2-RFA learned with respect to the uniform distribution as long as the maximum average noise rate over the attributes is less than 1/2 and there is at least one known relevant attribute with no noise.

Our results are based on the observation of Kalai *et al.* [1] that the so-called low-degree Fourier algorithm is a weak agnostic learner. In particular, our basic learning algorithm is a combination of the low-degree algorithm with a randomized algorithm due to Blum *et al.* [4]) that improves on the error bound of the basic low-degree algorithm. The proof of the algorithm's error bound also depends critically on a Fourier property of linear threshold functions over the cube due to Gotsman and Linial [3]. This basic algorithm provides the agnostic, attribute, and malicious noise results. The RCSAN algorithm adds on top of this basic algorithm some Fourier-based machinery for eliminating certain noise elements that the basic algorithm does not handle especially well.

Finally, we show that in relation to our learning algorithm for the standard attribute-noise model, the RCSAN model produces noise effects that are similar to those that can be produced by the malicious model. Potentially, then, the RCSAN model could be an interesting intermediary between the attribute noise model and the more difficult malicious model in other contexts as well.

2 Preliminaries

2.1 Fourier Transform

Many of our results make use of Fourier notation and basic results. For any function $f : \{0,1\}^n \rightarrow \mathbf{R}$ and for all $a \in \{0,1\}^n$, we define $\hat{f}(a) \equiv E_{x \sim \mathcal{U}_n}[f(x)\chi_a(x)]$, where $\mathcal{U}_n$ denotes the uniform distribution over $\{0,1\}^n$, $\chi_a(x) \equiv (-1)^{a \cdot x}$, and $a \cdot x$ represents the dot product of the bit vectors a and x. Each $\hat{f}(a)$ is a *Fourier coefficient* of f. The *Fourier representation* (or *expansion*) of f is $\sum_a \hat{f}(a)\chi_a$ and is equivalent to f. $\hat{f}(0^n)$ (0^n denotes the n-bit vector containing only 0's) is called the *constant Fourier coefficient.* The *first-order Fourier coefficients* are those coefficients for which $|a| = 1$, that is, for which a contains a single 1 bit.

We use e_i to denote the n-bit vector that has a single 1 in position i (bit locations are assumed to be numbered 1 through n). For two n-bit vectors a and b, $a \oplus b$ denotes the bitwise exclusive OR of the vectors. In particular, if $i \neq j$ then $e_i \oplus e_j$ represents the vector with 1's only in positions i and j.

In this paper, Boolean functions map to $\{-1,1\}$. Parseval's identity says that for any f, $E_{\mathcal{U}}[f^2] = \sum_a \hat{f}^2(a)$. This implies that if f is Boolean then $\sum_a \hat{f}^2(a) = 1$. It is easily seen that for all Boolean f and for all $a \in \{0,1\}^n$, $\hat{f}(a) = 2\Pr_{\mathcal{U}_n}[f = \chi_a] - 1 = 1 - 2\Pr_{\mathcal{U}_n}[f \neq \chi_a]$.

This paper focuses on noise-tolerant learnability of the class $\mathcal{L}$ of linearly separable functions over the Boolean cube. In Fourier terms, $\mathcal{L} = \cup_{n\geq 0}\mathcal{L}_n$, where $\mathcal{L}_n = \{\ell : \{0,1\}^n \to \{-1,1\} \mid \exists F = \sum_{|a|\leq 1} \hat{F}(a)\chi_a \text{ s.t. } \ell = \text{sign}(F)\}$.

2.2 Learning Models

The underlying learning model for this paper is PAC learning [5] with respect to the uniform distribution (or *with respect to uniform* for short); we assume that the reader is familiar with this model. We will often state that certain results hold "with high probability"; this should be understood to mean that these results hold with probability $1-\delta$ for arbitrary PAC confidence parameter $\delta > 0$. In this paper, algorithms will be considered efficient if they run in polynomial time in the number of inputs n, in an estimation tolerance parameter τ (bounds on which will in turn depend on parameters of the noise model), and in $\log(1/\delta)$.

With one exception, each noise model considered can be thought of as defining a noisy oracle that, on each query, first draws a noiseless example from a standard PAC example oracle $EX(f,\mathcal{U}_n)$ and then applies some noise process to this example, returning the resulting (possibly noisy) example as the response to the query. A *noiseless example* of a function f consists of a pair $(x, f(x))$, where x is called an *instance* (or input) and $f(x)$ is called the *label* (or output) of the example. The bits of an instance x are sometimes called the *attributes* of the instance. The notation (x^j, f^j) is used to represent the jth example returned by an oracle (either noiseless or noisy). If the example comes from a noisy oracle then—depending on the noise model—either or both of x^j and f^j may be noisy versions of an underlying noiseless example.

The *agnostic learning* model introduced by Kearns *et al.* [6] is the one exception mentioned above. It can be thought of as a particularly strong form of noise applied to the labels of examples, that is, as a form of *classification noise.* When learning $\mathcal{L}$ with respect to the uniform distribution, the strong version of this model becomes the following: the learner has access to an oracle $EX(f,\mathcal{U}_n)$ for an arbitrary Boolean function $f : \{0,1\}^n \to \{-1,1\}$. Given $\epsilon > 0$, the goal of the learner is to output a (possibly randomized) hypothesis $h : \{0,1\}^n \to \{-1,1\}$ such that $\Pr_{x\sim\mathcal{U}_n}[f(x) \neq h(x)] \leq \text{opt} + \epsilon$, where opt is the minimum of $\Pr_{\mathcal{U}_n}[\ell \neq f(x)]$ over all $\ell \in \mathcal{L}_n$. Here and elsewhere, in addition to the probability being over the uniform choice of x, it is also implicitly over the random choices made by h, if h is randomized (as it will be for our algorithms). Kearns *et al.* also consider a weak version of agnostic learning, wherein the goal is to find a *weak approximator* h to the target f (*i.e.*, h such that $\Pr_{\mathcal{U}_n}[h \neq f] \leq 1/2 - 1/p$ for some p polynomial in the learning parameters), given that f is weakly approximable by some function in $\mathcal{L}$.

In all of the other noise models considered, our goal will be to produce a hypothesis h that weakly approximates f with respect to uniform. In particular, we will say that $\mathcal{L}$ is *ϕ-learnable* for ϕ a function of the tolerance τ mentioned above and various parameters of the noise processes if there is a learning algorithm $\mathcal{A}$ that, given a noisy oracle for any $f \in \mathcal{L}$, produces (with high probability) a hypothesis h such that $\Pr_{\mathcal{U}_n}[h \neq f] \leq \phi$.

In the *attribute noise* model introduced by Shackelford and Volper [7], a noise distribution $\mathcal{N}$ over $\{0,1\}^n$ defines the behavior of the noise oracle $EX^{\mathcal{N}}(f,\mathcal{U})$. After drawing a noiseless example $(x, f(x))$, the attribute noise oracle draws $a \sim \mathcal{N}$ and returns as its output the noisy example $(x \oplus a, f(x))$.

In the *malicious noise* model introduced by Valiant [8], we will think of the noisy oracle *marking* each noiseless example $(x, f(x))$ with probability η. If an noiseless example is not marked, then it is returned as the oracle's output. Otherwise, the oracle is allowed to return an arbitrary, maliciously-chosen noisy example. The oracle can be assumed to be computationally unbounded, to know the target f, and even to know the current state of the learning algorithm.

The primary remaining noise model considered, the *restricted context-sensitive attribute noise* (RCSAN) model, will be described in a later section.

2.3 Restricted Focus of Attention

In the *Restricted Focus of Attention* (k-RFA) learning model introduced by Ben-David and Dichterman [2], the learner is only allowed to see k bits of each instance. The learner chooses the bits to be seen. The primary learning algorithm presented in this paper uses examples only to estimate the constant and first-order Fourier coefficients (over noisy examples). It is easy to see from the definition of these coefficients that they can all be estimated to inverse-polynomial accuracy given a polynomially large set of examples in the 1-RFA model. One version of RCSAN learning also needs to compute estimates of $E[\chi_{e_i}(x)\chi_{e_j}(x)]$ over noisy examples; this can clearly be accomplished in the 2-RFA model. Thus, all of our results apply in the 1-RFA or 2-RFA models, but in the sequel we will present the algorithms as if they are operating without any restriction on focus.

3 Weak Agnostic/Adversarial Noise Learning

In this section, we will show that $\mathcal{L}$ is weakly agnostically learnable with respect to the uniform distribution by a 1-RFA learner as long as the target f is such that there is some $\ell \in \mathcal{L}$ satisfying (roughly) $\Pr_{\mathcal{U}}[\ell \neq f] < 1/8$. However, we will find it convenient to first develop a learning result in a closely related noise model and return later to how this relates to weak agnostic learning. In the uniform-distribution *adversarial noise* model, after a target function $f \in \mathcal{L}$ has been selected but before learning begins, for some fixed $\eta > 0$ (the *adversarial noise rate*) an adversary is allowed to choose an arbitrary set of instances and corrupt their labels, producing a noisy Boolean function f^η that we will refer to as the η-*corrupted* version of f. The only limitation on f^η is that it must satisfy $\Pr_{\mathcal{U}_n}[f^\eta \neq f] \leq \eta$.

Theorem 1. *For any $\eta, \tau > 0$, $\mathcal{L}$ is efficiently* 1-*RFA* $(2\eta + \tau + 1/4)$-*learnable with respect to the uniform distribution despite adversarial noise of rate η.*

Proof. Fix any η and τ, let $f \in \mathcal{L}$, and let f^η be any adversarially η-corrupted version of f. Also assume that the PAC confidence parameter $\delta > 0$ is specified.

Our learning algorithm $\mathcal{A}$ will begin by drawing a set of $m = 25(n+1)^2 \ln(2(n+1)/\delta)/2\tau^2$ examples (x^j, f^j) from the noisy example oracle $EX(f^\eta, \mathcal{U})$. For each $|a| \leq 1$, $\mathcal{A}$ will then calculate $\hat{g}(a) \equiv (1/m)\sum_j f^j \chi_a(x^j)$. That is, for each such a, $\hat{g}(a)$ is an estimate of the Fourier coefficient $\widehat{f^\eta}(a)$ of the noisy function. Standard Hoeffding bounds [9] show that, with probability at least $1-\delta$ over the choice of examples, every $\hat{g}(a)$ will be within an additive factor of $\tau/2.5(n+1)$ of the corresponding $\widehat{f^\eta}(a)$. Next, from these estimated coefficients $\mathcal{A}$ constructs the (non-Boolean) function

$$g \equiv \sum_{|a|\leq 1} \hat{g}(a)\chi_a \ .$$

Finally, $\mathcal{A}$ defines the randomized Boolean function h as follows: $h(x) = -1$ with probability $p \equiv (1-g(x))^2/2(1+g^2(x))$ and $h(x) = 1$ with probability $1-p$. $\mathcal{A}$ outputs h as its hypothesis.

Clearly we can convert $\mathcal{A}$ to a 1-RFA algorithm by drawing a separate sample to compute each $\hat{g}(a)$, and both this RFA algorithm and the original are efficient. What remains to be shown is that for h as given above, with high probability $\Pr[h \neq f] \leq 2\eta + \tau + 1/4$.

The algorithm's definition of randomized Boolean h in terms of deterministic non-Boolean approximator g comes from Blum et al. [4], who show (in their Lemma 3) that for such an h and for any Boolean function f, $\Pr[h \neq f] \leq E[(f-g)^2]/2$. Furthermore, by Parseval's identity and the linearity of the Fourier transform, $E_{\mathcal{U}}[(f-g)^2] = \sum_a (\hat{f}(a) - \hat{g}(a))^2$. Since by the definition of g we have that $\hat{g}(a) = 0$ for all $|a| > 1$, breaking this sum into two parts gives us

$$\Pr_{\mathcal{U}}[h \neq f] \leq \frac{1}{2}\sum_{|a|\leq 1}\left(\hat{f}(a) - \hat{g}(a)\right)^2 + \frac{1}{2}\sum_{|a|>1}\hat{f}^2(a) \ . \tag{1}$$

Gotsman and Linial [3] have shown that for any $f \in \mathcal{L}$, $\sum_{|a|>1}\hat{f}^2(a) \leq 1/2$. Thus, what remains is to upper bound the first term of (1) by $2\eta+\tau$. The proof of this bound is similar to the proof of Observation 3 in [1] but uses an observation of Bshouty (personal communication) to achieve an improved 2η term rather than the 4η that would result from using the "almost triangle" inequality as in [1].

First, let $\alpha \equiv \tau/2.5(n+1)$ and recall that $\mathcal{A}$ chooses a sufficiently large set of examples such that, with high probability, for all $|a| \leq 1$ we have that $|\hat{g}(a) - \widehat{f^\eta}(a)| \leq \alpha$. This means that

$$\begin{aligned}
\sum_{|a|\leq 1}\left(\hat{f}(a) - \hat{g}(a)\right)^2 &\leq \sum_{|a|\leq 1}\left(|\hat{f}(a) - \widehat{f^\eta}(a)| + \alpha\right)^2 \\
&\leq \sum_{|a|\leq 1}\left(\hat{f}(a) - \widehat{f^\eta}(a)\right)^2 + 2\sum_{|a|\leq 1}|\hat{f}(a) - \widehat{f^\eta}(a)|\alpha + \sum_{|a|\leq 1}\alpha^2 \\
&\leq \sum_{|a|\leq 1}\left(\hat{f}(a) - \widehat{f^\eta}(a)\right)^2 + 5\sum_{|a|\leq 1}\alpha
\end{aligned}$$

since the Fourier coefficients of the Boolean functions f and f^η all fall in the range $[-1,1]$. Thus the first term in (1) is bounded by $(1/2)\sum_{|a|\leq 1}(\hat{f}(a)-\widehat{f^\eta}(a))^2+\tau$. Furthermore,

$$\begin{aligned}\sum_{|a|\leq 1}\left(\hat{f}(a)-\widehat{f^\eta}(a)\right)^2 &\leq \sum_{a\in\{0,1\}^n}\left(\hat{f}(a)-\widehat{f^\eta}(a)\right)^2\\ &= E\left[(f-f^\eta)^2\right]\\ &= 4\Pr[f\neq f^\eta]\\ &\leq 4\eta\end{aligned}$$

where the first equality follows by again applying Parseval's identity and the second because f and f^η are both $\{-1,1\}$-valued. □

In agnostic learning terms, what we have shown is that if the target f is such that there exists an $\ell\in\mathcal{L}$ and a $\gamma>\tau$ satisfying $\Pr_{\mathcal{U}}[f\neq\ell]\leq 1/8-\gamma/2$ then algorithm $\mathcal{A}$ above will (with high probability) output a randomized hypothesis h such that $\Pr_{\mathcal{U}}[h\neq f]\leq 1/2-(\gamma-\tau)$, which for sufficiently large $\gamma-\tau$ means that h weakly approximates f. Thus, algorithm $\mathcal{A}$ in fact 1-RFA weakly agnostically learns $\mathcal{L}$ with respect to uniform.

4 Attribute Noise

Bshouty *et al.* [10] showed that the class AC^0 of polynomial-size constant-depth AND/OR circuits can be learned despite certain types of attribute noise. In particular, given mild constraints on ϵ and δ, if the attribute noise is defined by a known product distribution in which the noise rate for each bit is at most inverse polylogarithmic in n then AC^0 is learnable with respect to the uniform distribution despite such attribute noise. Based on their analysis and the observations above, we will next show that $\mathcal{L}$ is weakly learnable with respect to uniform despite an unknown attribute noise process, subject to only the mildest of constraints.

Specifically, we will make use of the following easily-shown observation from Bshouty *et al.* (part of the proof of their Theorem 8):

Lemma 1 (Bshouty et al.). *Let $\mathcal{N}$ be any noise distribution over $\{0,1\}^n$ and let $f:\{0,1\}^n\to\{-1,1\}$ be any Boolean function. Then for each $c\in\{0,1\}^n$, $E_{x\sim U_n,a\sim\mathcal{N}}[f(x)\chi_c(x\oplus a)]=\hat{f}(c)E_{a\sim\mathcal{N}}[\chi_c(a)]$.*

For the linear Fourier coefficients $\hat{f}(e_i)$, note that

$$E_{a\sim\mathcal{N}}[\chi_{e_i}(a)]=E_{a\sim\mathcal{N}}[(-1)^{a_i}]=1-2\Pr_{a\sim\mathcal{N}}[a_i=1]\ .$$

Thus, for any attribute noise distribution $\mathcal{N}$ and Boolean function f, given a set S of examples $\{(x^j,f^j)\}$ generated by the attribute-noise oracle $EX^{\mathcal{N}}(f,\mathcal{U})$, the expected value of $(1/|S|)\sum_S f^j \chi_{e_i}(x^j)$ is $\hat{f}(e_i)(1 - 2\Pr_{a\sim\mathcal{N}}[a_i = 1])$. We write $\widehat{f^{\mathcal{N}}}(e_i)$ to denote this expected value.

For any given noise distribution $\mathcal{N}$ define $p_{\mathcal{N}} \equiv \max_{1\le i\le n}\Pr_{a\sim\mathcal{N}}[a_i = 1]$. That is, $p_{\mathcal{N}}$ is an upper bound on the marginal error rate of each of the attributes. We will show that $\mathcal{L}$ can be weakly learned with respect to uniform despite unknown attribute noise $\mathcal{N}$, where $\mathcal{N}$ is arbitrary except for the constraint that $p_{\mathcal{N}}$ must be non-negligibly less than 1/2 in order to achieve weak learning efficiently. Since learning is information-theoretically impossible given uniform attribute noise of rate 1/2 (as this in effect replaces each instance with some other uniform-random instance), this is a very weak constraint on the noise process.

Theorem 2. *For any $\tau > 0$ and any unknown distribution $\mathcal{N}$ over $\{0,1\}^n$ such that $p_{\mathcal{N}} < 1/2$, $\mathcal{L}$ is efficiently* 1*-RFA* $(p^2_{\mathcal{N}} + \tau + 1/4)$*-learnable with respect to uniform despite unknown $\mathcal{N}$-attribute noise.*

Proof. The proof is very similar to that of Theorem 1, except that the algorithm $\mathcal{A}$ described in that proof now operates on examples generated by an $EX^{\mathcal{N}}(f,\mathcal{U})$ oracle rather than by an $EX(f^{\eta},\mathcal{U})$ oracle. Specifically, $\mathcal{A}$ will use the $\mathcal{N}$ oracle to estimate, for all $|a| \le 1$, $\hat{g}(a)$'s that are approximations to the coefficients $\widehat{f^{\mathcal{N}}}(a) = \hat{f}(a)(1 - 2\Pr_{a\sim\mathcal{N}}[a_i = 1])$ to a tolerance of $\tau/2.5(n+1)$. The function g is defined in terms of these estimated coefficients as before, and h is again defined in terms of g. From the proof of Theorem 1 we have that

$$\Pr[h \ne f] \le \frac{1}{2}\sum_{|a|\le 1}(\hat{f}(a) - \widehat{f^{\mathcal{N}}}(a))^2 + \tau + \frac{1}{2}\sum_{|a|>1}\hat{f}^2(a) \ .$$

Since we are considering attribute noise only, $\hat{f}(0^n) = \widehat{f^{\mathcal{N}}}(0^n)$. For every $|a| = 1$, by the definition of $p_{\mathcal{N}}$, $(\hat{f}(a) - \widehat{f^{\mathcal{N}}}(a))^2 \le 4p^2_{\mathcal{N}}\hat{f}^2(a)$. So $\sum_{|a|\le 1}(\hat{f}(a) - \widehat{f^{\mathcal{N}}}(a))^2 \le 4p^2_{\mathcal{N}}\sum_{|a|\le 1}\hat{f}^2(a) = 4p^2_{\mathcal{N}} - 4p^2_{\mathcal{N}}\sum_{|a|>1}\hat{f}^2(a)$, where the equality follows from Parseval's identity. Inserting this into bound on $\Pr[h \ne f]$ above gives

$$\Pr[h \ne f] \le 2p^2_{\mathcal{N}} + \tau + \left(\frac{1}{2} - 2p^2_{\mathcal{N}}\right)\sum_{|a|>1}\hat{f}^2(a) \ .$$

Since our assumed constraint on $p_{\mathcal{N}}$ implies that $1/2 > 2p^2_{\mathcal{N}}$, this bound is maximized when $\sum_{|a|>1}\hat{f}^2(a)$ is maximized. Using the fact that $\sum_{|a|>1}\hat{f}^2(a) \le 1/2$ completes the proof. □

5 Malicious Noise

Recall that in the malicious noise model, conceptually each example is "marked" independently with probability η, and those that are marked can be corrupted arbitrarily by a malicious adversary. In this model, the worst case for the algorithm $\mathcal{A}$ of Theorem 1—in terms of the bound we can prove on the approximation error of $\mathcal{A}$'s hypothesis h relative to the target f—is when the adversary chooses to make every marked example identical to all the other marked examples. This approach can be used to maximize the difference that can be achieved for a given set of marked examples between $\mathcal{A}$'s estimated coefficients $\{\hat{g}(a) : |a| \leq 1\}$ and the corresponding true coefficients $\hat{f}(a)$, which in turn maximally increases (weakens) the bound on $\Pr[h \neq f]$ provided by (1) over the $1/4 + \tau$ bound that would apply in the noise-free setting.

The magnitude of the error induced by this worst-case malicious noise process in the estimate of a fixed first-order coefficient $\hat{f}(e_i)$ depends on the magnitude of the coefficient. For instance, if the coefficient value is 0 (that is, the attribute i is irrelevant) then on average the adversary will only change the value of attribute i in half of the marked examples; the other half will already have the desired attribute value. On the other hand, if $|\hat{f}(e_i)| = 1$ then attribute i will be changed in every marked example, and the magnitude of the expected difference between $\mathcal{A}$'s estimate of $\hat{f}(e_i)$ and the true value will be 2η. The error induced in the estimate of $\hat{f}(0^n)$ similarly depends on the magnitude of this coefficient.

It follows that, for fixed marking rate η and estimation tolerance $\tau > 0$, applying algorithm $\mathcal{A}$ of the proof of Theorem 1 to malicious noise examples will with high probability produce Fourier estimates $\hat{g}(a)$ such that $(1/2)\sum_{|a|\leq 1}(\hat{f}^2(a) - \hat{g}^2(a)) \leq 4(n+1)\eta^2 + \tau$. Thus, the algorithm without modification will weakly learn $\mathcal{L}$ despite malicious noise of rate $\eta = \Omega(1/\sqrt{n})$.

However, it would obviously be a simple matter to modify the algorithm to detect a large number of identical examples and, once detected, to ignore them in computing the coefficients $\hat{g}(a)$. In fact, notice that a set of such examples corrupted in this way would no longer be uniformly distributed over the instance space, and in particular notice that the attributes would no longer be independent.

Comparing the attribute and malicious noise models, then, there are (at least) two key differences. First, while the attribute noise model adds an error vector to an underlying instance, the malicious noise model replaces the underlying instance in its entirety. Second, as Bshouty *et al.* [10] point out, uniformly distributed instances remain uniform after arbitrary attribute noise is applied, while (as we have just seen) this is not necessarily the case with malicious noise. That said, the malicious noise model does allow the adversary to consider the entire example when corrupting an individual attribute, so the adversary can potentially craft the corrupted examples so that the overall set of examples still appears to be drawn uniformly.

This comparison of models suggests that it might be worthwhile to consider noise models that lie between the attribute and malicious models. We consider this direction in the next section.

6 Context-Sensitive Attribute Noise

In the *restricted context-sensitive attribute noise* (RCSAN) model, the noise process is similar to that of attribute noise, but the process is potentially sensitive to the label and a limited number of attribute values. This is somewhat analogous to the RFA learning model, except it is the noise process that is restricted here. The specific version of the model considered here could be called 1-RCSAN, since we will allow the noise applied to an attribute i of an example to depend only on the example label and on the value of i itself. In the sequel, we will simply call this the RCSAN model.

Each instantiation of the RCSAN model defines four noise rates p_i^{++}, p_i^{+-}, p_i^{-+}, and p_i^{--} for each attribute $1 \leq i \leq n$. If a given pre-noise example $(x, f(x))$ is such that $\chi_{e_i}(x) = +1$ (that is, $x_i = 0$) and $f(x) = -1$ then the noise process will flip x_i from 0 to 1 with probability p_i^{+-}. The other three noise rates similarly define the probability of attribute i being corrupted in the remaining three attribute/label contexts.

This model generalizes the product attribute noise model, in which each attribute i is assigned a single context-free noise rate p_i that is applied to attribute i in every example, regardless of the value of the attribute or the label. As we saw earlier, when a uniform-distribution learning algorithm is based on estimates of first-order Fourier coefficients, the general attribute noise model—in which an arbitrary (possibly non-product) noise distribution $\mathcal{N}$ is allowed—effectively reduces to a form of product attribute noise. So, for algorithms based on estimating first-order Fourier coefficients, the restricted context-sensitive attribute noise model is strictly stronger than the attribute noise model considered in section 4.

Furthermore, with respect to the type of error induced in Fourier coefficients, the RCSAN model is in some ways more similar to malicious noise than to attribute noise. In particular, recall that the errors induced by the attribute noise model in the first-order Fourier coefficients of a target function are multiplicative in nature: each coefficient is reduced by a multiplicative factor as small as $1-2p_{\mathcal{N}}$. On the other hand, like the malicious noise model, the RCSAN model can induce additive error in the first-order Fourier coefficients. For example, consider an irrelevant attribute i, that is, an attribute for which $\hat{f}(e_i) = 0$. If this coefficient is estimated as the sample mean of $f^j \chi_{e_i}(x^j)$ over a set of noisy examples $\{(x^j, f^j)\}$ where the noise rates are $p_i^{++} = p_i^{--} = 0$ and $p_i^{+-} = p_i^{-+} = \eta > 0$, then the expected value of the estimate will be η.

In the remainder of this section, we will examine uniform-distribution RCSAN-tolerant learning of a subclass of linear threshold functions, the class $\mathcal{L}_h$ of *homogeneous linear threshold functions*. This class is the discrete analog of the origin-centered halfspaces considered by Kalai *et al.* [1] and others. Specifically, $\mathcal{L}_h$ is the set of all functions $f : \{0,1\}^n \rightarrow \{-1,1\}$ such that there is a function $F = \sum_{|a|=1} \hat{F}(a)\chi_a$ and $f = \text{sign}(F)$. We'll begin with several simple lemmas showing that $\mathcal{L}_h$ has a number of nice Fourier properties.

6.1 Properties of $\mathcal{L}_h$

Lemma 2. *If $f \in \mathcal{L}_h$ then f is balanced, that is, $E_{x \sim \mathcal{U}_n}[f(x)] = \hat{f}(0^n) = 0$.*

Proof. Let $\bar{x}$ represent the bitwise-complement of $x \in \{0,1\}^n$. Since $f \in \mathcal{L}_h$, there is some F such that for every $x \in \{0,1\}^n$, $f(x) = \text{sign}(\sum_{|a|=1} \hat{F}(a)\chi_a(x))$.

Fixing such an F we have that for all x, $f(\bar{x}) = \text{sign}(\sum_{|a|=1} \hat{F}(a)\chi_a(\bar{x})) = \text{sign}(-\sum_{|a|=1} \hat{F}(a)\chi_a(x)) = -f(x)$. It follows that $E_{x\sim\mathcal{U}_n}[f(x)] = 0$. □

Lemma 3. *If $f : \{0,1\}^n \to \{-1,1\}$ is in $\mathcal{L}_h$ and $1 \leq i \leq n$ then for any $b \in \{-1,1\}$, $\Pr_{x\sim\mathcal{U}_n}[f(x) = 1 \wedge \chi_{e_i}(x) = b] = \Pr_{x\sim\mathcal{U}_n}[f(x) = -1 \wedge \chi_{e_i}(x) = -b]$.*

Proof. Fix arbitrary $b \in \{-1,1\}$. By the proof of the preceding lemma, we know that for all $x \in \{0,1\}^n$, $f(\bar{x}) = -f(x)$. Thus, for every $x \in \{0,1\}^n$ such that $f(x) = 1$ and $\chi_{e_i}(x) = b$ there is a distinct $y = \bar{x}$ such that $f(y) = -1$ and $\chi_{e_i}(y) = -b$. Therefore, the set of such x's is no larger than the set of such y's. But it is similarly easy to see that the set of such y's is no larger than the set of such x's. Thus the sets are of equal size and have equal probability with respect to the uniform distribution. □

Lemma 4. *If $f : \{0,1\}^n \to \{-1,1\}$ is in $\mathcal{L}_h$ and $1 \leq i \leq n$ then for any $b \in \{-1,1\}$,*

$$\Pr_{x\sim\mathcal{U}_n}[f(x) = 1 \wedge \chi_{e_i}(x) = b] = \frac{1 + b\hat{f}(e_i)}{4}.$$

Proof. By the definition of Fourier coefficients and the previous lemma, $\hat{f}(e_i) = 2\Pr[f = \chi_{e_i}] - 1 = 2(\Pr[f = \chi_{e_i} = 1] + \Pr[f = \chi_{e_i} = -1]) - 1 = 4\Pr[f = \chi_{e_i} = 1] - 1$. This proves the $b = 1$ case. The $b = -1$ case can be proved similarly by starting with $\hat{f}(e_1) = 1 - 2\Pr[f \neq \chi_{e_i}]$. □

Lemma 5. *If $f : \{0,1\}^n \to \{-1,1\}$ is in $\mathcal{L}_h$ and $1 \leq i \neq j \leq n$ then for any $b_1, b_2 \in \{-1,1\}$, $\Pr_{x\sim\mathcal{U}_n}[f(x) = 1 \wedge \chi_{e_i}(x) = b_1 \wedge \chi_{e_j}(x) = b_2] = \Pr_{x\sim\mathcal{U}_n}[f(x) = -1 \wedge \chi_{e_i}(x) = -b_1 \wedge \chi_{e_j}(x) = -b_2]$*

Proof. The proof is essentially the same as that of Lemma 3. □

Lemma 6. *If $f : \{0,1\}^n \to \{-1,1\}$ is in $\mathcal{L}_h$ and $1 \leq i \neq j \leq n$ then $\hat{f}(e_i \oplus e_j) = 0$.*

Proof. Let χ_{ij} represent $\chi_{e_i\oplus e_j}$ and define χ_i and χ_j similarly in terms of e_i and e_j, respectively. Then applying the definition of Fourier coefficients and the preceding lemma, we have that

$$\begin{aligned}
\hat{f}(e_i \oplus e_j) &= 2\Pr_{x\sim\mathcal{U}_n}[f = \chi_{ij}] - 1 \\
&= 2(\Pr[f = 1 \wedge \chi_i = 1 \wedge \chi_j = 1] + \Pr[f = 1 \wedge \chi_i = -1 \wedge \chi_j = -1] + \\
&\quad \Pr[f = -1 \wedge \chi_i = 1 \wedge \chi_j = -1] + \Pr[f = -1 \wedge \chi_i = -1 \wedge \chi_j = 1]) - 1 \\
&= 2(\Pr[f = 1 \wedge \chi_i = 1 \wedge \chi_j = 1] + \Pr[f = 1 \wedge \chi_i = -1 \wedge \chi_j = -1] + \\
&\quad \Pr[f = 1 \wedge \chi_i = -1 \wedge \chi_j = 1] + \Pr[f = 1 \wedge \chi_i = 1 \wedge \chi_j = -1]) - 1 \\
&= 2\Pr[f = 1] - 1.
\end{aligned}$$

Since f is balanced (by Lemma 2), $\Pr[f = 1] = 1/2$. □

Lemma 7. *If* $f : \{0,1\}^n \to \{-1,1\}$ *is in* $\mathcal{L}_h$ *and* $1 \le i \ne j \le n$ *then for any* $b_1, b_2 \in \{-1,1\}$,

$$\Pr_{x \sim \mathcal{U}_n}[f(x) = 1 \wedge \chi_{e_i}(x) = b_1 \wedge \chi_{e_j} = b_2] = \frac{1 + b_1 \hat{f}(e_i) + b_2 \hat{f}(e_j)}{8}.$$

Proof. Fix b_1, b_2, i, and j and let f_{ij} represent the projection of f that ignores attributes i and j and instead treats every example x as if these attributes have constant values such that $\chi_{e_i}(x) = b_1$ and $\chi_{e_j}(x) = b_2$. It follows from the Fourier representation of f that $E[f_{ij}] = \hat{f}(0^n) + b_1\hat{f}(e_i) + b_2\hat{f}(e_j) + b_1 b_2 \hat{f}(e_i \oplus e_j)$. Furthermore, based on Lemma 2 and the preceding lemma, we know that this sum reduces to $b_1\hat{f}(e_i) + b_2\hat{f}(e_j)$. Of course, $E[f_{ij}]$ is also equal to $2\Pr[f_{ij} = 1] - 1 = 2\Pr[f = 1 \mid \chi_{e_i} = b_1 \wedge \chi_{e_j} = b_2] - 1$. Applying the definition of conditional probability and solving for $\Pr[f = 1 \wedge \chi_{e_i} = b_1 \wedge \chi_{e_j} = b_2]$ gives the lemma. □

6.2 Learning $\mathcal{L}_h$

With these lemmas in hand, let us now consider the effect of context-sensitive noise on the estimate of a first-order Fourier coefficient of a homogeneous linear threshold function.

Lemma 8. *Let* $f : \{0,1\}^n \to \{-1,1\}$ *be any function in* $\mathcal{L}_h$ *and let* $1 \le i \le n$. *Then for any RCSAN process, the expected value of the sample mean of* $f^j \chi_{e_i}(x^j)$ *over a set of noisy examples* $\{(x^j, f^j)\}$ *is*

$$\frac{p_i^{+-} + p_i^{-+} - p_i^{++} - p_i^{--}}{2} + \hat{f}(e_i)\left(1 - \frac{p_i^{+-} + p_i^{-+} + p_i^{++} + p_i^{--}}{2}\right) \qquad (2)$$

Proof. The expected value without noise is of course $\hat{f}(e_i)$. By Lemma 4, the probability that $f = \chi_{e_i} = 1$—which is also the probability that noise rate p_i^{++} applies—is $(1 + \hat{f}(e_i))/4$. The effect of attribute noise on these examples is to subtract 1 rather than adding 1 to $\sum_j f^j \chi_{e_i}(x^j)$. Thus, the expected effect of noise due to examples where $f = \chi_{e_i} = 1$ is to add $-p_i^{++}(1 + \hat{f}(e_i))/2$ to the true expected value $\hat{f}(e_i)$. Similarly, applying Lemma 3 as well as Lemma 4, the expected contribution of noise due to examples where $f = \chi_{e_i} = -1$ is $-p_i^{--}(1 + \hat{f}(e_i))/2$. Further applications of Lemmas 3 and 4 to the remaining cases gives that the expected value of the sample mean is

$$\hat{f}(e_i) + \frac{(p_i^{+-} + p_i^{-+})(1 - \hat{f}(e_i)) - (p_i^{++} + p_i^{--})(1 + \hat{f}(e_i))}{2}.$$

Rearranging this expression gives the lemma. □

Thus, in general, the noise induced in a coefficient $\hat{f}(e_i)$ by an RCSAN process is a combination of additive error (of rate $(p_i^{+-} + p_i^{-+} - p_i^{++} - p_i^{--})/2$) and multiplicative error (of rate $1 - (p_i^{+-} + p_i^{-+} + p_i^{++} + p_i^{--})/2$).

Obviously, if the RCSAN process generating noisy examples is known, then this theorem can be used to recover a close approximation to the noiseless Fourier coefficient $\hat{f}(e_i)$ from the noisy estimate of this coefficient as long as the multiplicative factor in (2) is bounded away from 0, or equivalently, as long as the average noise rate $\bar{p}_i \equiv (p_i^{+-} + p_i^{-+} + p_i^{++} + p_i^{--})/4$ is bounded away from 1/2. So it is easy to learn $\mathcal{L}_h$ in the RCSAN model if the noise process is known and does not completely obscure the target function.

The more interesting case, then, is if the noise process is unknown but perhaps constrained in some way. For instance, consider the constraint that for all i, the average noise probability when f and χ_{e_i} agree ($(p_i^{++} + p_i^{--})/2$) is equal to the average noise probability when they disagree. Then the additive term in (2) will vanish. In this situation, it can be seen that Theorem 2 applies, with the modification that we will use $\bar{p} \equiv \max_{i=1..n} \bar{p}_i$ in place of $p_{\mathcal{N}}$. In fact, if the additive term in (2) is nonzero but less than, say, $\tau/5(n+1)$ for all i, then we can modify $\mathcal{A}$ to use a (polynomial) sample size m' such that the $\hat{g}(a)$'s computed are all (with high probability) within $\tau/5(n+1)$ of the true mean values they estimate. The result is that (with high probability) each $\hat{g}(a)$ will be within $\tau/2.5(n+1)$ of its mean value, as needed for the remainder of the proof of Theorem 2. In short, as long as for every attribute i the average noise rate $\bar{p}_i$ is non-negligibly less than 1/2 and the differences $(p_i^{+-} + p_i^{-+}) - (p_i^{++} + p_i^{--})$ are all sufficiently small, then Theorem 2 applies and $\mathcal{L}_h$ is weakly learnable with respect to uniform despite an unknown RCSAN process.

This is of course a very strong constraint on the RCSAN process. The main result of this section shows how to learn $\mathcal{L}_h$ with a much milder constraint on the RCSAN process.

Theorem 3. *For any $\tau > 0$ and given any RCSAN process, $\mathcal{L}_h$ is efficiently 2-RFA $(\bar{p}^2 + \tau + 1/4)$-learnable with respect to uniform. The RCSAN process is unknown and unconstrained except that $\bar{p} < 1/2$, there must be one known attribute k for which $\bar{p}_k = 0$, and there must be a known non-negligible value $\beta > 0$ such that $|\hat{f}(e_k)| > \beta$.*

Proof. (Sketch) The key is showing that, for every attribute $i \neq k$, we can obtain a good approximation to the additive error $((p_i^{+-} + p_i^{-+}) - (p_i^{++} + p_i^{--}))/2$ present in $\hat{g}(e_i)$ computed as the mean value of $f^j \chi_i(x^j)$ over a set of RCSAN examples $\{(x^j, f^j)\}$ (where as before χ_i is shorthand for χ_{e_i}). Once this additive error has been (mostly) eliminated from the $\hat{g}(e_i)$'s, the analysis above applies, and we can use a slight modification of the algorithm of Theorem 2 to obtain our result. So we will show how to estimate the additive error.

Let E_{ik} represent the expected value of $\chi_i(x^j)\chi_k(x^j)$ over random noisy examples (x^j, f^j) drawn according to some fixed RCSAN process. Note that, since attribute k is assumed to be noise free, if attribute i is also noise free then $E_{ik} = E_{\mathcal{U}}[\chi_i \chi_k] = 0$. Now consider how this changes if $p_i^{++} > 0$. By Lemma 7 we know that with probability $(1 + \hat{f}(e_i) + \hat{f}(e_k))/8$ a pre-noise example x is such that $f(x) = \chi_i(x) = \chi_k(x) = 1$. Since corrupting bit i of such an x changes $\chi_i(x)\chi_k(x)$ from $+1$ to -1, the net change in E_{ik} due to positive p_i^{++} over these x's is $-p_i^{++}(1+\hat{f}(e_i)+\hat{f}(e_k))/4$. On the other hand, with probability $(1+\hat{f}(e_i)-\hat{f}(e_k))/8$ we have $f(x)=\chi_i(x)=1$ and $\chi_k(x)=-1$, and the net change in E_{ik} due to positive p_i^{++} over these x's is $p_i^{++}(1+\hat{f}(e_i)-\hat{f}(e_k))/4$. Combining these effects, the overall change in E_{ik} due to positive p_i^{++} will be $-p_i^{++}\hat{f}(e_k)/2$. Applying Lemma 5 along with Lemma 7, we can similarly see that the contribution to E_{ik} due to positive p_i^{--} will be $-p_i^{--}\hat{f}(e_k)/2$. On the other hand, the total change due to positive p_i^{+-} and p_i^{-+} will be $(p_i^{+-}+p_i^{-+})\hat{f}(e_k)/2$. Overall, then, we see that $E_{ik}=\hat{f}(e_k)((p_i^{+-}+p_i^{-+})-(p_i^{++}+p_i^{--}))/2$.

Our estimate for the additive error term in $\hat{g}(e_i)$, then, will be obtained by drawing a noisy sample, computing sample means that approximate E_{ik} and $\hat{f}(e_k)$, and dividing the approximation of E_{ik} by the approximation of $\hat{f}(e_k)$. We will use a sample size large enough so that this quotient is, with high probability, within an additive factor of $\tau/5(n+1)$ of the expected additive error term in (2). Based on the earlier discussion, it should be clear that such an estimate will be sufficiently close to give us the learning result claimed.

Specifically, we will use a sample large enough to guarantee with high probability that the noisy estimate of E_{ik} is additively within $O(\beta\tau/n)$ of its expected value. By standard Hoeffding bounds, a polynomial number of examples will suffice. We will then estimate (with high probability) $\hat{f}(e_k)$ to within a multiplicative factor c close enough to 1 to achieve the desired bound on the additive error in the quotient of our estimates. It can be shown that $|1-c|=O(\beta\tau/n)$ is sufficient for this purpose, and Chernoff bounds tell us that the sample size required will again be polynomial. □

7 Further Work

An obvious question whenever uniform-distribution weak learning results are derived is how far the results can be extended beyond uniform. The extant proofs of the results underlying Gotsman-Linial's observation seem to rely heavily on independence and other properties of the uniform distribution, so such a generalization may not be easy. However, if the results could be extended to a sufficiently general set of distributions, this might lead to noise-tolerant uniform-distribution strong learning algorithms for $\mathcal{L}$.

There may be interesting subclasses of $\mathcal{L}$ such that for any function f in the class the constant and first-order Fourier coefficients represent much more than half of the power spectrum of f. If the spectral power of the low-order coefficients of all of the functions in such a class were over 3/4, then results of Kalai *et al.* [1] could be applied to give an efficient algorithm weakly agnostically learning $\mathcal{L}$ using $\mathcal{L}$ as the hypothesis class. Do such subclasses of $\mathcal{L}$ exist? The class of Majority functions is not such a subclass, as it can be shown that asymptotically the low-order coefficients for odd Majority functions represent roughly $2/\pi \approx .64$ of the power spectrum. Alternatively, can the Kalai *et al.* results be strengthened so that they could be applied to weaker approximators?

The fact that $\mathcal{L}$ can be weakly learned despite an essentially optimal rate of adversarial noise can be shown to imply that the constant 2 in the bound of

Theorem 1 cannot be improved unless the bound is also changed in some other way. How tight is the bound of Theorem 1?

Kalai *et al.* [1] also explore malicious noise learning and give a simple algorithm for uniformly learning halfspaces over the unit sphere that tolerates noise rate η up to roughly $\Omega(1/n^{1/4})$. It would be nice to have a comparable result over the cube (although it may require unrestricted focus of attention).

Can an RCSAN result similar to Theorem 3 be obtained without the need for a known noise-free attribute? Beyond this, it may be interesting to explore 1-RCSAN learnability of other classes as well as k-RCSAN learning of $\mathcal{L}$ and other classes for $k > 1$.

Acknowledgements

Rocco Servedio posed a question that started this work and reminded the author of the Gotsman-Linial result. Nati Linial graciously explained a part of the proof of that result. Nader Bshouty pointed out an improvement to the proof of the Kalai *et al.* [1] observation that the low-degree algorithm is a weak agnostic learner. Discussions with Rocco and Adam Klivans were also very helpful.

References

1. Kalai, A.T., Klivans, A.R., Mansour, Y., Servedio, R.A.: Agnostically learning halfspaces. In: Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science. (2005) 11–20
2. Ben-David, S., Dichterman, E.: Learning with restricted focus of attention. In: Proceedings of the 6th Annual Conference on Computational Learning Theory. (1993) 287–296
3. Gotsman, C., Linial, N.: Spectral properties of threshold functions. Combinatorica **14** (1994) 35–50
4. Blum, A., Furst, M., Jackson, J., Kearns, M., Mansour, Y., Rudich, S.: Weakly learning DNF and characterizing statistical query learning using Fourier analysis. In: Proceedings of the 26th Annual ACM Symposium on Theory of Computing. (1994) 253–262 Preliminary version available as `http://www.mathcs.duq.edu/~jackson/dnfsq.ps`.
5. Valiant, L.G.: A theory of the learnable. Communications of the ACM **27** (1984) 1134–1142
6. Kearns, M.J., Schapire, R.E., Sellie, L.M.: Toward efficient agnostic learning. Machine Learning **17** (1994) 115–141
7. Shackelford, G., Volper, D.: Learning k-DNF with noise in the attributes. In: Proceedings of the 1988 Workshop on Computational Learning Theory. (1988) 97–103
8. Valiant, L.G.: Learning disjunctions of conjunctions. In: Proceedings of the Ninth International Joint Conference on Artificial Intelligence. Volume 1. (1985) 560–566
9. Hoeffding, W.: Probability inequalities for sums of bounded random variables. American Statistical Association Journal **58** (1963) 13–30
10. Bshouty, N.H., Jackson, J.C., Tamon, C.: Uniform-distribution attribute noise learnability. Inf. Comput. **187** (2003) 277–290