Electronic Colloquium on Computational Complexity, Report No. 88 (2005)

Learning Juntas in the Presence of Noise

Jan Arpe[*]

Institut für Theoretische Informatik, Universität zu Lübeck
Ratzeburger Allee 160, 23538 Lübeck, Germany
Email: arpe@tcs.uni-luebeck.de

August 3, 2005

Abstract

The combination of two major challenges in machine learning is investigated: dealing with large amounts of irrelevant information and learning from noisy data. It is shown that large classes of Boolean concepts that depend on a small number of variables—so-called *juntas*—can be learned efficiently from random examples corrupted by random attribute and classification noise.

To accomplish this goal, a two-phase algorithm is presented that copes with several problems arising from the presence of noise: firstly, a suitable method for approximating Fourier coefficients in the presence of noise is applied to infer the relevant variables. Secondly, as one cannot simply read off a truth table from the examples as in the noise-free case, an alternative method to build a hypothesis is established and applied to the examples restricted to the relevant variables.

In particular, for the class of monotone juntas depending on d out of n variables, the sample complexity is polynomial in $\log(n/\delta)$, 2^d, γ_a^{-d}, and γ_b^{-1}, where δ is the confidence parameter and $\gamma_a, \gamma_b > 0$ are noise parameters bounding the noise rates away from $1/2$. The running time is bounded by the sample complexity times a polynomial in n.

So far, all results hold for the case of uniformly distributed examples—the only case that (apart from side notes) has been studied in the literature yet. We show how to extend our methods to non-uniformly distributed examples and derive new results for monotone juntas.

For the attribute noise, we have to assume that it is generated by a product distribution since otherwise fault-tolerant learning is in general impossible: we construct a noise distribution P and a concept class $\mathcal{C}$ such that it is impossible to learn $\mathcal{C}$ under P-noise.

Keywords: learning of Boolean functions, noise-tolerant learning, learning under irrelevant information, juntas, Fourier analysis

[*] Supported by DFG research grant Re 672/4.

ISSN 1433-8092

1 Introduction

Learning in the presence of huge amounts of irrelevant information and learning in the presence of noise have attracted considerable interest in the past. In this paper, we investigate what happens when these worlds collide: How can we learn Boolean concepts that depend on only a small number d of attributes—so-called *d-juntas*—under the unpleasant effects of attribute and classification noise?

Efficient learning in the presence of irrelevant information is considered to be among the most important and challenging issues in machine learning (see Mossel, O'Donnell, and Servedio [22]) with a wide range of applications (see Akutsu, Miyano, and Kuhara [1] and Blum and Langley [7]). The goal is to design fast algorithms that learn from a number of examples that may depend exponentially on d (since the output hypotheses are represented by their truth tables being of size 2^d) but only logarithmically on the number n of all attributes. While this goal has been achieved for various junta subclasses and learning models (see e.g. Littlestone [19]), it is an open question whether the class of all n-ary d-juntas can be PAC-learned efficiently under the uniform distribution. The fastest algorithm to date was proposed by Mossel et al. [22] and runs in time $n^{0.704d} \cdot \mathrm{poly}(n, 2^d, \log(1/\delta))$, where δ is the confidence parameter. Their algorithm combines two methods: the *Fourier method* infers relevant variables via estimating Fourier coefficients and the *parity method* learns the concept via solving linear equations over GF(2). In particular, the Fourier method yields an algorithm for learning the class of monotone d-juntas in time $\mathrm{poly}(n, 2^d, \log(1/\delta))$.

As coping with irrelevant information has been identified as a core challenge in many machine learning applications, it is most natural to take into account that real-world data are often disturbed by noise. Angluin and Laird [2] were the first to investigate PAC-learning in the presence of classification noise, whereas attribute-noise was first considered for the class of k-DNF formulas by Shackelford and Volper [24] and later by Decatur and Gennaro [10]. Bshouty, Jackson, and Tamon [9] introduced the notion of *noisy distance* between concepts and showed how this quantity relates to uniform-distribution PAC-learning in the presence of attribute and classification noise. Further aspects of learning in noisy settings were investigated by Goldman and Sloan [13] and by Miyata, Tarui, and Tomita [20].

Our main contribution is an algorithm that efficiently learns large classes of juntas despite the presence of almost arbitrary attribute and classification noise. Thus we manage to cope with both problems: irrelevant information and noise.

More precisely, we assume that a learning algorithm receives uniformly distributed examples $(x_1, \ldots, x_n, y) \in \{0,1\}^n \times \{-1,+1\}$ in which each attribute value x_i is flipped independently with probability p_i and the sign of the label y is switched with probability η. To avoid that the noise-affected data is turned into completely random noise, we require that there be constants $\gamma_a, \gamma_b > 0$ such that for all attribute noise rates p_i, $|1 - 2p_i| \geq \gamma_a$ and for the classification noise rate η, $|1 - 2\eta| \geq \gamma_b$. We call such noise distributions *(γ_a, γ_b)-bounded noise.* We show that the class of Boolean functions we call *s-low d-juntas* is exactly learnable from $\mathrm{poly}(\log n, 2^d, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$ examples in time $n^s \cdot \mathrm{poly}(n, 2^d, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$ under (γ_a, γ_b)-bounded noise. Roughly speaking, a concept is s-low if it suffices to check all Fourier coefficients up to the s-th level in order to find all relevant attributes (see Section 3). As a main application, the class of monotone d-juntas, for which $s = 1$, is learnable in time $\mathrm{poly}(n, 2^d, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$ under (γ_a, γ_b)-bounded noise.

A function is 1-low iff its function value is correlated with each relevant variable, i.e., for each relevant variable x_i, at least one of the subfunctions $f_{x_i=0}$ and $f_{x_i=1}$ is unbalanced. Coincidentally, it has been shown that certain greedy algorithms studied by Arpe and Reischuk [3, 4] and Fukagawa and Akutsu [11] successfully infer all relevant attributes of such functions in case of uniformly distributed attributes disturbed by attribute noise with small noise rates. Compared to the greedy method, the Fourier technique can cope with almost arbitrary attribute noise rates. While the Fourier method easily extends to concepts of a higher degree of balance, such an extension is not known for the greedy method.

We now briefly describe how we solve the manifold problems that occur when trying to extend results from the noise-free case to the noisy case. In the noise-free setting, it is trivial to achieve the time bound $n^d \cdot \mathrm{poly}(n, 2^d, \log(1/\delta))$ for the whole class of n-ary d-juntas by testing for all subsets of d variables whether these are relevant. This is accomplished by checking whether the examples restricted to these variables do not contain any contradictions. In the noisy case, however, there is no obvious way to check whether a subset of the variables is relevant. We solve this problem by adapting the Fourier method presented by Mossel et al. [22]. For this it is necessary to approximate Fourier coefficients of Boolean functions from highly disturbed data.

Also, in the noise-free setting, once the relevant variables are inferred, one can just read off a truth table from the undisturbed examples. This is impossible in case of unreliable data. To overcome this problem, we apply a learning algorithm for arbitrary concepts to the examples restricted to the relevant variables. This restriction is essential since in this way, the number of examples needed to build a hypothesis does not depend on n but only on d. The learning algorithm uses the Fourier-based learning approach originated by Linial, Mansour, and Nisan [17] and extended to the noisy scenario by Bshouty, Jackson, and Tamon [9]. A direct application of the algorithm of Bshouty et al. yields a sample complexity of $n^{d+O(1)}$. By first applying our procedure to detect all relevant attributes, we significantly improve this sample complexity to depend only polylogarithmically on n (and exponentially on d).

So far all results are valid for uniformly distributed attribute vectors—the only case for which positive noise-tolerant learning results have previously been obtained in the literature (as far as we are aware). We extend our methods to non-uniform attribute distributions, i.e., the oracle first draws an example according to a product distribution D with rates $d_1, \ldots, d_n \in [\gamma_c, 1 - \gamma_c]$ for some $\gamma_c > 0$ and then applies (γ_a, γ_b)-bounded noise. We show that in this setting, monotone d-juntas are learnable from $\mathrm{poly}(\log n, 2^{d^2}, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$ examples in time $\mathrm{poly}(n, 2^{d^2}, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$, provided that $\gamma_c \geq 0.2764$. It turns out that the extension is not as straightforward as one might first think: while the method for the case of uniformly distributed attributes relies on the fact that the orthonormal basis of parity functions is compatible with the *exclusive or* operation used in the noise model, this is no more the case for the biased orthonormal bases that are appropriate for non-uniform distributions. We solve this problem by combining *unbiased* parity functions with *biased* inner products. As a consequence, the analysis becomes a lot more intricate since in order to approximate a biased Fourier coefficient $\hat{f}(I)$, $I \subseteq [n]$, one already has to have good approximations to all coefficients $\hat{f}(J)$, $J \subsetneq I$. In addition, we have to provide a lower bound on the absolute value of nonzero biased Fourier coefficients for monotone juntas (see Section 5 for details).

Finally, we prove that without restricting the attribute noise distributions (for example to product distributions), noise-tolerant learning is in general impossible: we construct an attribute noise distribution P (that is not a product distribution) and a concept class $\mathcal{C}$ such that it is impossible to learn $\mathcal{C}$ under P-noise. In particular, this shows that our results may not be extended to arbitrary noise distributions.

Our proofs have three main ingredients: standard Hoeffding bounds [14], harmonic analysis of Boolean functions under uniform [6] and non-uniform [12] distribution, and the *noise operator* T_P, a generalization of the *Bonami-Beckner operator* T_ρ, formally introduced by O'Donnell [23] and previously studied in several contexts [15, 5, 21, 9].

In Section 2, we introduce basic notation, definitions, and tools and present the considered learning and noise model. After reviewing how to learn juntas in the noise-free case in Section 3, we show how to handle the noisy case in Section 4. Section 5 deals with the extension of our results to non-uniformly distributed attributes.

2 Preliminaries

We consider Boolean functions $f : \{0,1\}^n \to \{-1,+1\}$, also called *concepts*. The class of all n-variate concepts is denoted by $\mathcal{B}^n$. A concept is *monotone* if for all $x, y \in \{0,1\}^n$ such that $x \le y$, we have $f(x) \ge f(y)$ (note that for variables, the value 1 for "true" is larger than the value 0 for "false", whereas for function values -1 (true) and 1 (false), it is the other way round). For $I \subseteq [n] = \{1, \dots n\}$, we define the *parity function* $\chi_I \in \mathcal{B}^n$ by $\chi_I(x) = (-1)^{\sum_{i \in I} x_i}$. For $x, y \in \{0,1\}^n$, $x \oplus y$ denotes the vector obtained from component-wise exclusive or. We denote probabilities by $\mathbb{P}$ and expectations by $\mathbb{E}$. The uniform distribution over $\{0,1\}^n$ is denoted by U_n. The functions log and ln denote the binary and the natural logarithm, respectively.

A *concept class* is a set of concepts $f \in \mathcal{B}^n$. Let $\mathcal{C}$ be a concept class and $f \in \mathcal{C}$. A vector $(x_1, \dots, x_n, y) \in \{0,1\}^n \times \{-1,+1\}$ is called an *example*. It is *consistent with* f if $f(x_1, \dots, x_n) = y$. A sequence of m examples is called a *sample of size* m.

Consider the space $\mathbb{R}^{\{0,1\}^n}$ of real-valued functions on the hypercube. The inner product $\langle f, g \rangle = \mathbb{E}_{x \sim U_n}[f(x)g(x)]$ induces the norm $\|f\|_2 = \sqrt{\langle f, f \rangle}$ and turns $\mathbb{R}^{\{0,1\}^n}$ into a Hilbert space of dimension 2^n with orthonormal basis $(\chi_I \mid I \subseteq [n])$, see for example Bernasconi [6].

Let $f : \{0,1\}^n \to \mathbb{R}$ and $I \subseteq [n]$. The *Fourier coefficient of* f *at* I is

$$\hat{f}(I) = \mathbb{E}_{x \sim U_n}[f(x) \cdot \chi_I(x)] = 2^{-n} \textstyle\sum_{x \in \{0,1\}^n} f(x) \cdot \chi_I(x) \; .$$

If $I = \{i\}$, we write $\hat{f}(i)$ instead of $\hat{f}(\{i\})$. We have the *Fourier expansion*

$$f(x) = \textstyle\sum_{I \subseteq [n]} \hat{f}(I) \cdot \chi_I(x) \tag{1}$$

for all $x \in \{0,1\}^n$. Given a sample $S = (x^k, y^k)_{k\in[m]} \in (\{0,1\}^n \times \{-1,+1\})^m$ (with $y^k = f(x^k)$), define the *empirical Fourier coefficient of f at I given S* by

$$\tilde{f}_S(I) = \tfrac{1}{m}\textstyle\sum_{k=1}^m \chi_I(x^k)\cdot y^k \ . \tag{2}$$

By standard Hoeffding bounds [14], $\tilde{f}_S(I)$ approximates $\hat{f}(I)$ up to an additive error of ε with probability at least $1-\delta$, provided that $m \geq 2\cdot\ln(\delta/2)\cdot(1/\varepsilon^2)$ uniformly distributed examples are given.

A function $f \in \mathcal{B}^n$ *depends* on variable x_i (and x_i is *relevant* to f) if the $(n-1)$-variate subfunctions $f_{x_i=0}$ and $f_{x_i=1}$ with x_i set to 0 and 1, respectively, are not equal. Denote the set of relevant variables of f by $\mathrm{rel}(f)$. A function that depends on at most d variables is called a *d-junta*, and the class of n-variate Boolean d-juntas is denoted by $\mathcal{J}_d^n$. The class of monotone d-juntas is denoted by MON_d^n, and the class of juntas such that the function restricted to its relevant variables is symmetric is denoted by SYM_d^n.

To learn a *target concept* $f \in \mathcal{C}$, we assume that a learning algorithm has access to a noisy *example oracle $EX_{P,\eta}(f)$*, where $P : \{0,1\}^n \to [0,1]$ is a probability distribution called the *attribute noise distribution* and $\eta \in [0,1]$ is the *classification noise rate*. On request, $EX_{P,\eta}(f)$ first generates an attribute vector $x \in \{0,1\}^n$ according to U_n and computes $y = f(x)$. Then it generates an *attribute noise vector* $a \in \{0,1\}^n$ according to P and a *classification noise bit* $b \in \{-1,+1\}$ which is set to -1 with probability η and to 1 with probability $1-\eta$. Finally it returns the *(P,η)-noisy example* $(x \oplus a, y \cdot b)$. If an example oracle applies only attribute noise, we denote it by $EX_{P,-}(f)$. If no noise is applied at all, we just write $EX(f)$. Let $\delta \in (0,1]$ be a *confidence parameter*. An algorithm $\mathcal{A}$ *exactly learns* the class $\mathcal{C}$ under noise (P,η) (or *(P,η)-learns C*) with *confidence* $1-\delta$ if for any target concept $f \in \mathcal{C}$, given access to $EX_{P,\eta}(f)$, $\mathcal{A}$ outputs a *hypothesis* $h \in \mathcal{B}^n$ such that with probability at least $1-\delta$, $h = f$. The class $\mathcal{C}$ is *exactly (P,η)-learnable* if there is an algorithm $\mathcal{A}$ that on any input $\delta > 0$, learns $\mathcal{C}$ under noise (P,η) with confidence $1-\delta$. The number of calls to $EX_{P,\eta}(f)$ is called the *sample complexity* of $\mathcal{A}$.

For the time being, we restrict ourselves to uniformly distributed attribute values. The case of non-uniform distributions is discussed in Section 5.

Since arbitrary attribute noise distributions often turn out to make learning impossible, we often restrict ourselves to *product attribute noise* considered for example by Goldman and Sloan [13]. Here, each attribute x_i of an example is flipped independently with some probability $p_i \in [0,1]$, called the *(attribute) noise rate* of x_i. Thus we have $P(a_1,\ldots,a_n) = \prod_{i=1}^n p_i^{a_i}\cdot(1-p_i)^{1-a_i}$.

3 Learning Juntas—A Review of the Noise-Free Case

In this section we review the "Fourier algorithm" described by Mossel et al. [22]. We first look at how one can learn monotone juntas and then show how to extend the method to learn larger subclasses of juntas. This will be helpful to make clear why we are interested in *s-low juntas* and to understand the methods presented in Section 4.

Let $f \in \mathrm{MON}_d^n$ be a monotone d-junta. It is well known (cf. [22]) that f is correlated with all of its relevant variables, i.e., the probability that x_i and $f(x)$ take the same value differs from 1/2 and thus $\hat{f}(i) = \mathbb{P}_{x \sim U_n}[f(x) = x_i] - \mathbb{P}_{x \sim U_n}[f(x) \neq x_i] \neq 0$. This fact may be exploited to infer the relevant variables of f from (uniformly distributed) random examples $(x^k, f(x^k))$, $x^k \in \{0,1\}^n$, $k \in [m]$, as follows: simply approximate the Fourier coefficients $\hat{f}(i)$ by the empirical coefficients $\tilde{f}(i)$ defined in (2). If sufficiently many independent examples are available, then with high probability, the relevant variables are exactly those for which $\tilde{f}(i)$ is sufficiently far away from zero, i.e., $|\hat{f}(i)| \geq \tau$ for some threshold $\tau > 0$.

Once we have correctly inferred the relevant variables, it is easy to derive a consistent hypothesis: we obtain an appropriate truth table by restricting the given examples to the relevant variables. With high probability (see Blumer et al. [8]), there is only one hypothesis having the same set of relevant variables and being consistent with the function table, namely the target concept f.

Clearly, the approach also works for non-monotone functions with the property that all relevant variables are correlated with the function value. Moreover, we can use the following fact (implicitly used in Mossel et al. [22]) to extend the method to larger classes of Boolean functions:

Lemma 3.1 *Let $f \in \mathcal{B}^n$. Then for all $i \in [n]$, x_i is relevant to f if and only if there exists $I \subseteq [n]$ such that $i \in I$ and $\hat{f}(I) \neq 0$.*

Hence whenever we find a nonzero Fourier coefficient $\hat{f}(I)$, we know that all variables x_i, $i \in I$, are relevant to f. Moreover, all relevant variables can be detected in this way, and we only have to check out subsets of size at most $d = |\operatorname{rel}(f)|$. However, there are $\Theta(n^d)$ such subsets, an amount that we would like to reduce. This leads us to the following definition:

Definition 3.2 Let $f \in \mathcal{J}_d^n$, $x_i \in \operatorname{rel}(f)$, and $s \in [d]$. Variable x_i is *s-low for f* if there exists an $I \subseteq [n]$ such that $i \in I$, $|I| \leq s$, and $\hat{f}(I) \neq 0$. The concept f is *s-low* if all $x_i \in \operatorname{rel}(f)$ are s-low for f. The set of s-low d-juntas is denoted by $\mathcal{R}_d^n(s)$.

In these terms, monotone juntas are 1-low, i.e., $\mathrm{MON}_d^n \subseteq \mathcal{R}_d^n(1)$. Even more: all juntas that are locally (anti-)monotone are 1-low; these are juntas that can be turned into a monotone function by negating some input variables. This includes all monomials and clauses of arbitrary literals. Actually, the vast majority of juntas belongs to $\mathcal{R}_d^n(1)$ since a random junta fulfills $\hat{f}(i) \neq 0$ for all $x_i \in \operatorname{rel}(f)$ with overwhelming probability, see Blum and Langley [7] and Mossel et al. [22].

Also for other subclasses $\mathcal{C}$ of $\mathcal{J}_d^n$, finding the smallest s such that $\mathcal{C} \subseteq \mathcal{R}_d^n(s)$ has recently attracted considerable interest: The class of all unbalanced d-juntas is contained in $\mathcal{R}_d^n((2/3) \cdot d)$ (see Mossel et al. [22]), and the class $\mathrm{SYM}_d^n \setminus \{\chi_I \mid |I| \leq d\}$ of symmetric d-juntas that are not parity functions is now known to be contained in $\mathcal{R}_d^n(O(d/\log d))$ (see Kolountzakis et al. [16]).

In the left part of Fig. 1, we present the algorithm (which we call `IRV`) described by Mossel et al. [22] for inferring the relevant variables of s-low d-juntas.

Proposition 3.3 ([22]) *Let $f \in \mathcal{R}_d^n(s)$ be an s-low d-junta. Then with probability at least $1 - \delta$, algorithm `IRV` exactly infers the relevant variables of f from a sample* *of size* $\mathrm{poly}(\log n, 2^d, \log(1/\delta))$ *in time* $n^s \cdot \mathrm{poly}(n, 2^d, \log(1/\delta))$.

4 Learning Juntas—The Noisy Case

Now let us see what we can do if the example generating oracle behaves unreliably. We first introduce the noise operator T_P which is crucial to the proofs of Section 4.2. In addition, this operator adds some structure and insight to results of Bshouty et al. [9].

4.1 The Noise Operator

For an attribute noise distribution $P : \{0,1\}^n \to [0,1]$, we define the *noise operator* $T_P : \mathbb{R}^{\{0,1\}^n} \to \mathbb{R}^{\{0,1\}^n}$ by $T_P(f)(x) = \mathbb{E}_{a\sim P}[f(x \oplus a)]$. If $f \in \mathcal{B}^n$, then $T_P(f)(x)$ is the expected value returned by the oracle $EX_{P,-}(f)$ provided that x is the outcome of the oracle's draw of the attribute vector according to U_n. By linearity of the mean, T_P is a linear operator.

For $I \subseteq [n]$ and $a \sim P$, let p_I be the probability that an odd number of bits a_i with $i \in I$ is set to one, i.e.,

$$p_I = \mathbb{P}_{a\sim P}[\chi_I(a) = -1] \tag{3}$$

and let $\alpha_I = \mathbb{E}_{a\sim P}[\chi_I(a)] = 1 - 2p_I$. The following lemma states how the Fourier coefficients of $T_P(f)$ are related to those of f:

Lemma 4.1 *Let* $f : \{0,1\}^n \to \mathbb{R}$, P *be an attribute noise distribution, and* $I \subseteq [n]$. *Then* $\widehat{T_P(f)}(I) = \alpha_I \cdot \hat{f}(I)$.

Proof. We have

$$\begin{aligned}\widehat{T_P(f)}(I) &= \mathbb{E}_{x\sim U_n}[T_P(f)(x) \cdot \chi_I(x)] \;=\; \mathbb{E}_{x\sim U_n}[\mathbb{E}_{a\sim P}[f(x \oplus a) \cdot \chi_I(x)]] \\ &= \mathbb{E}_{a\sim P}[\mathbb{E}_{x\sim U_n}[f(x \oplus a) \cdot \chi_I(x)]] \;=\; \mathbb{E}_{a\sim P}[\mathbb{E}_{x\sim U_n}[f(x) \cdot \chi_I(x \oplus a)]] \\ &= \mathbb{E}_{a\sim P}[\mathbb{E}_{x\sim U_n}[f(x) \cdot \chi_I(x) \cdot \chi_I(a)]] \;=\; \alpha_I \cdot \hat{f}(I)\ .\end{aligned}$$

□

Several additional properties of T_P used in the proofs of Section 4.2 are provided by the following lemma.

Lemma 4.2 *Let* $f : \{0,1\}^n \to [-1,1]$ *and* $P : \{0,1\}^n \to [0,1]$ *be an attribute noise distribution. Then*

(a) $T_P(f)(x) = \sum_{I\subseteq[n]} \alpha_I \hat{f}(I)\chi_I$ *for all* $x \in \{0,1\}^n$.

(b) $\|T_P(f)\|_1 = \mathbb{E}_{x\sim U_n}[\,|\,\mathbb{E}_{a\sim P}[f(x \oplus a)]\,|\,]$.

(c) $\|T_P(f)\|_2^2 = \mathbb{E}_{x\sim U_n}[\,(\mathbb{E}_{a\sim P}[f(x \oplus a)])^2\,] = \sum_{I\subseteq[n]} \alpha_I^2 \hat{f}(I)^2$.

(d) $\|T_P(f)\|_2^2 \le \|T_P(f)\|_1 \le \|T_P(f)\|_2$.

(e) $\|T_P(f)\|_2^2 \ge \min_{I\subseteq[n]} \alpha_I^2 \cdot \|f\|_2^2$.

Proof. Part (a) follows by Fourier expansion (1), part (b) is immediate from the definitions, and part (c) follows from the definitions and from Parseval's equation $\|f\|_2^2 = \sum_{I\subseteq[n]} \hat{f}(I)^2$. The first inequality of part (d) follows since for all $g : \{0,1\}^n \to [-1,+1]$, we have

$$\|g\|_2^2 = \sum_{x\in\{0,1\}^n} g(x)^2 \le \sum_{x\in\{0,1\}^n} |g(x)| = \|g\|_1 \,.$$

Clearly, $|T_P(f)(x)| \le 1$ for all $x \in \{0,1\}^n$ if $|f(x)| \le 1$ for all $x \in \{0,1\}^n$. The second inequality of part (d) follows from $\mathbb{E}[|X|]^2 \le \mathbb{E}[X^2]$ for real-valued random variables X. Finally, part (e) is an immediate consequence of part (c). □

From these properties, it follows that $\|T_P(f-g)\|_1$ equals twice the *noisy distance* between f and g introduced by Bshouty et al. [9]. Furthermore, one of their main results, which is also used in our proofs, easily follows from Lemma 4.2:

Theorem 4.3 ([9]) *Let $P : \{0,1\}^n \to [0,1]$ be a probability distribution and $f, g \in \mathcal{B}^n$. Then $\frac{1}{2}\|T_P(f-g)\|_2^2 \le \|T_P(f-g)\|_1 \le \|T_P(f-g)\|_2$.*

4.2 Approximating Fourier Coefficients

Given a uniformly distributed (P,η)-noisy sample, the empirical Fourier coefficient $\tilde{f}_S(I)$ approximates $\mathbb{E}_{x\sim U_n, a\sim P, b\sim\eta}[\chi_I(x\oplus a)\cdot f(x)\cdot b)]$. It is easy to see (cf. Bshouty et al. [9]) that this expectation equals $(1-2p_I)\cdot(1-2\eta)\cdot\hat{f}(I)$ with p_I as defined in (3). Using standard Hoeffding bounds [14], we obtain

Lemma 4.4 *Let $f \in \mathcal{B}^n$, P be an attribute noise distribution and $\eta \in [0,1]$ be a classification noise rate. Let $\delta, \varepsilon > 0$ and S be a (P,η)-noisy sample of size $m \ge 2\cdot\ln(2/\delta)\cdot(1/\varepsilon^2)$. Then $|\tilde{f}_S(I) - (1-2p_I)(1-2\eta)\hat{f}(I)| \le \varepsilon$ with probability at least $1-\delta$.*

Thus we can infer $\hat{f}(I)$ from $\tilde{f}(I)$ by this method if and only if $p_I \ne 1/2$. Unfortunately, it can happen that $p_I = 1/2$ for some I (even if $\mathbb{P}_{a\sim P}[a_i = -1] \ne 1/2$ for all $i \in [n]$). Even worse, we can construct a concept class $\mathcal{C}$ and an attribute noise distribution P such that $\mathcal{C}$ is (information-theoretically) not $(P,-)$-learnable:

Theorem 4.5 *There is a concept class $\mathcal{C}$ and an attribute noise distribution P such that $\mathcal{C}$ is not $(P,-)$-learnable. In addition, P may be chosen such that $p_i < 1/2$ for all $i \in [n]$.*

Proof. We set $n = 2$, $P(00) = 4/8$, $P(01) = 3/8$, $P(10) = 1/8$, and $P(11) = 0$. Then $\mathbb{P}[a_1 = 1] = 1/8$ and $\mathbb{P}[a_2 = 1] = 3/8$. Let $f(x) = \chi_{[2]}(x) = (-1)^{x_1+x_2}$ and choose $\mathcal{C} = \{f, -f\}$. By Lemma 4.2 (c), $\|T_P(2f)\|_2^2 = 2\sum_{I\subseteq[n]} \alpha_I^2 \widehat{\chi_{[2]}}(I)^2 = 2\alpha_{[2]}^2 = (4/8 - 3/8 - 1/8 + 0)^2 = 0$ (note that $\hat{\chi}_I(J) = 1$ iff $I = J$ and $\hat{\chi}_I(J) = 0$ otherwise), hence also $\|T_P(2f)\|_1 = 0$. This implies that $|\mathbb{E}_{a\sim P}[f(x\oplus a) - (-f)(x\oplus a)]| = 0$ for all $x \in \{0,1\}^n$. But from this it follows that $(x, f(x\oplus a))$ and $(x, -f(x\oplus a))$ with $x \sim U_n$ and $a \sim P$ are identically distributed (see the proof of [9, Theorem 2]). Thus $(x\oplus a, f(x)b)$ and $(x\oplus a, -f(x)b)$ (with $\mathbb{P}[b = -1] = \eta$) are identically distributed by [9, Lemma 1]. Hence f and $-f$ are information-theoretically indistinguishable under P-attribute noise. □

In contrast, things look much nicer for product distributions P with noise rates p_i that are all different from $1/2$:

Definition 4.6 (γ_a-bounded product distribution) Let P be a product distribution with rates $p_1, \ldots, p_n$ and $\gamma_a > 0$. P is called a γ_a*-bounded product distribution* if for all $i \in [n]$, $|1 - 2p_i| \geq \gamma_a$.

It is easy to prove by induction that γ_a-bounded product distributions satisfy

$$\forall I \subseteq [n] : |1 - 2p_I| \geq \gamma_a^{|I|} \tag{4}$$

From now on, we restrict ourselves to γ_a-bounded product distributions. However, all results extend to arbitrary distributions for which condition (4) holds.

If all $p_I \neq 1/2$, then all Fourier coefficients are approximable, hence the whole target concept can be approximated via its Fourier expansion (1). Consequently, all concepts are learnable under these conditions by computing the hypothesis

$$h(x) = \operatorname{sgn} \textstyle\sum_{I \subseteq [n]} \frac{\tilde{f}(I)}{(1-2p_I)\cdot(1-2\eta)} \cdot \chi_I(x) \ . \tag{5}$$

The necessary sample and time complexity are as follows:

Proposition 4.7 *Let $\mathcal{C} = \mathcal{B}^n$, P be a γ_a-bounded product attribute noise distribution, and η be a classification noise rate such that $\gamma_b = |1 - 2\eta| > 0$. Then $\mathcal{C}$ is exactly (P, η)-learnable with confidence* $1 - \delta$ *using sample complexity and running time* $\mathrm{poly}(2^n, \log(1/\delta), \gamma_a^{-n}, \gamma_b^{-1})$.

Proof. For any $\varepsilon > 0$, Bshouty et al. [9] defined $\Delta_P^\varepsilon(\mathcal{C})$ to be the minimum noisy distance between ε-far concepts inside $\mathcal{C}$, i.e.,

$$\Delta_P^\varepsilon(\mathcal{C}) \ = \ \min \left\{ \tfrac{1}{2} \|T_P(f - g)\|_1 \ \middle| \ f, g \in \mathcal{C} : \ \tfrac{1}{2} \|f - g\|_1 > \varepsilon \right\} \ .$$

Thus $\Delta_P^\varepsilon(\mathcal{C})$ measures how close ε-far concepts in $\mathcal{C}$ can become when T_P is applied to them.

By [9, Theorem 8], choosing $\varepsilon = 2^{-n-1}$ and $T_\varepsilon = 2^{[n]}$, it remains to bound $\Delta_P^\varepsilon(\mathcal{C})$ from below to prove the claim. Note that PAC-learning with accuracy $1 - 2^{-n-1}$ is just exact learning since concepts differing in a fraction of inputs that is smaller than 2^{-n} must be equal. As observed in Section 4.2 (see (4)), $|\alpha_I| = |1 - 2p_I| \geq \gamma_a^{|I|}$.

Let $f, g \in \mathcal{C}$ be distinct concepts. Since $(f(x) - g(x))/2 \in \{-1, 0, +1\}$ for all $x \in \{0, 1\}^n$, we have $\|(f - g)/2\|_2^2 = \|(f - g)/2\|_1 \geq 2^{-n}$. By Lemma 4.2 (e), we have

$$\|T_P((f - g)/2)\|_2^2 \ \geq \ \min_{I \subseteq [n]} \alpha_I^2 \cdot \|(f - g)/2\|_2^2 \ \geq \ \gamma_a^{2n} \cdot 4 \cdot \|f - g\|_1 \ \geq \ \gamma_a^{2n} \cdot 2^{-n+2} \ .$$

By Theorem 4.3, $\frac{1}{2}\|T_P(f - g)\|_1 \geq 2^{-n}\gamma_a^{2n}$, yielding $\Delta_P^\varepsilon(\mathcal{C}) \geq 2^{-n}\gamma_a^{2n}$. Thus $1/\Delta_P^\varepsilon(\mathcal{C})$ is linear in 2^n and polynomial in γ_a^{-n}, and the desired result follows from [9, Theorem 8] which yields a sample and time complexity polynomial in $1/\varepsilon$, $\log(1/\delta)$, $1/(1 - 2\eta)$, $|T_\varepsilon|$ and $1/\Delta_P^\varepsilon(\mathcal{C})$. □

Although sample and time complexity are exponential in n, the method described will prove useful as part of our noise-tolerant learning algorithm for juntas (see Section 4.4).

Since d-juntas have all of their Fourier weight located in levels $0, \ldots, d$ (by Lemma 3.1), we obtain a better (but still not satisfactory) sample and time complexity by summing only over all $I \subseteq [n]$ of size at most d in equation (5).

Proposition 4.8 *Let P be a γ_a-bounded product attribute noise distribution and η be a classification noise rate such that $\gamma_b = |1 - 2\eta| > 0$. Then $\mathcal{J}_d^n$ is exactly (P, η)-learnable with confidence $1 - \delta$ using sample complexity and running time $n^d \cdot \mathrm{poly}(n, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$.*

Proof. We proceed similarly as in the proof of Proposition 4.7, but choose $\varepsilon = 2^{-d-1}$ and $T_\varepsilon = \{I \subseteq [n] \mid |I| \leq d\}$ (since $\hat{f}(I) = 0$ for all I of size larger than d). It remains to bound $\Delta_P^\varepsilon(\mathcal{J}_d^n)$ (as defined above in the proof of Proposition 4.7) from below. By (4), $|\alpha_I| = |1 - 2p_I| \geq \gamma_a^{|I|}$. Consequently, for distinct concepts $f, g \in \mathcal{J}_d^n$ and $h = f - g$, h depends on at most $2d$ variables, i.e., $\hat{h}(I) = 0$ whenever $|I| > 2d$. We have

$$\|T_P(h)\|_2^2 \geq \gamma_a^{4d} \cdot \sum_{I \subseteq [n]} \hat{h}(I)^2 \geq \gamma_a^{4d} \cdot 4\varepsilon = \gamma_a^{4d} \cdot 2^{-d+1} .$$

By Theorem 4.3, $\frac{1}{2}\|f - g\|_1 \geq 2^{-d-1} \cdot \gamma_a^{4d}$, yielding $\Delta_P^\varepsilon(\mathcal{J}_d^n) = \frac{1}{2}\|f - g\|_1 \geq 2^{-d-1} \cdot \gamma_a^{4d}$. Thus $1/\Delta_P^\varepsilon(\mathcal{J}_d^n)$ is linear in 2^d and polynomial in γ_a^{-d}, and the desired result follows from [9, Theorem 8]. □

Unfortunately, sample and time complexity do not drop for subclasses such as the monotone juntas since the Fourier weight may be spread evenly over all $\Theta(n^d)$ nonzero coefficients (as it is the case for example for monomials, see e.g. [23, Sec. 3.3]).

In the sequel we show how to combine the method just described with the idea of first detecting the relevant variables, as we did in the noise-free case. In Theorem 4.11, we show that this significantly reduces the sample complexity from $O(n^{d+O(1)})$ to $\mathrm{poly}(\log n, 2^d)$. In addition, for s-low d-juntas with $s < d$, also the running time decreases from $O(n^{d+O(1)})$ to $O(n^{s+O(1)})$.

4.3 Inferring the Relevant Variables

The detection of relevant variables works similarly to the noise-free case. The following modifications to the algorithm `IRV` (shown in the left part of Fig. 1) vaccinate it against noise; the resulting algorithm `Noisy-IRV` is shown in the right part of Fig. 1.

Firstly, the noisy version has to obtain some information about the noise parameters. In the variant presented here, it receives bounds $\gamma_a, \gamma_b > 0$ such that $|1 - 2p_i| \geq \gamma_a$ for all $i \in [n]$ and $|1 - 2\eta| \geq \gamma_b$ as additional inputs. Secondly, the number of examples that have to be drawn increases by a factor of $4 \cdot (\gamma_a^s \cdot \gamma_b)^{-2}$. Furthermore, the noise-free oracle $EX(f)$ is replaced by the noisy oracle $EX_{P,\eta}(f)$. In particular, in line 2 of `Noisy-IRV`, $x^k = x'^k \oplus a^k$ and $y^k = y'^k \cdot b^k$ for appropriate noise-free data x'^k, y'^k and noise a^k, b^k. Next, to ensure that in line 5 of the algorithm, β is an appropriate measure to decide whether the Fourier coefficient $\hat{f}(I)$

```
Algorithm IRV

input $\delta \in (0,1]$, $s \in [d]$
request $m = 2 \cdot \ln(\frac{2n}{\delta}) \cdot 2^{2d}$
  examples $(x^k, y^k)_{k \in [m]}$ from $EX(f)$
$R \leftarrow \emptyset$
for $I \subseteq [n]$ with $1 \leq |I| \leq s$ do
  $\beta \leftarrow \frac{1}{m} \cdot \sum_{k=1}^{m} \chi_I(x^k) \cdot y^k$
  if $|\beta| \geq 2^{-d-1}$
    then $R \leftarrow R \cup \{x_i \mid i \in I\}$
output ``relevant variables:'' $R$
```

```
Algorithm Noisy-IRV

input $\delta \in (0,1]$, $s \in [d]$, $\gamma_a, \gamma_b > 0$
request $m = 8 \cdot \ln(\frac{2n}{\delta}) \cdot 2^{2d} \cdot (\gamma_a^s \cdot \gamma_b)^{-2}$
  examples $(x^k, y^k)_{k \in [m]}$ from $EX_{P,\eta}(f)$
$R \leftarrow \emptyset$
for $I \subseteq [n]$ with $1 \leq |I| \leq s$ do
  $\beta \leftarrow (\gamma_a^{|I|} \cdot \gamma_b)^{-1} \cdot \frac{1}{m} \cdot \sum_{k=1}^{m} \chi_I(x^k) \cdot y^k$
  if $|\beta| \geq 2^{-d-1}$
    then $R \leftarrow R \cup \{x_i \mid i \in I\}$
output ``relevant variables:'' $R$
```

Figure 1: Algorithms IRV (Infer Relevant Variables) and Noisy-IRV to infer all relevant variables of concepts in $\mathcal{R}_d^n(s)$ in the noise-free and the noisy case, respectively.

vanishes, we divide the expression given in the noise-free setting by $\gamma_a^{|I|} \cdot \gamma_b$, which is a lower bound for $|1 - 2p_I| \cdot |1 - 2\eta|$.

Theorem 4.9 *Let $f \in \mathcal{R}_d^n(s)$ be an s-low junta. Let P be a γ_a-bounded attribute noise distribution and η be a classification noise rate such that $\gamma_b = |1 - 2\eta| > 0$. Then with probability $1 - \delta$, on input $\delta, s, \gamma_a, \gamma_b$, the variables classified as "relevant" by* Noisy-IRV *are exactly the relevant variables of f.*

Proof. Let S be the (P, η)-noisy sample that the algorithm obtains from the oracle $EX_{P,\eta}(f)$. Let $t = 2^{-d}$. Noisy-IRV classifies x_i as "relevant" if and only if $|\tilde{f}_S(I)| \geq (1/2) \cdot \gamma_a^{|I|} \cdot \gamma_b \cdot t$ for some I of size at most s with $i \in I$. By Lemma 4.4, for every $I \subseteq [n]$ of size at most s,

$$|\tilde{f}_S(I) - (1 - 2p_I)(1 - 2\eta)\hat{f}(I)| \leq \tfrac{1}{2} \cdot \gamma_a^s \cdot \gamma_b \cdot t \qquad (6)$$

with probability at least $1 - \frac{\delta}{n}$.

Consider some variable $x_i \in \mathrm{rel}(f)$. By assumption, there exists an $I \subseteq [n]$ of size at most s such that $i \in I$ and $\hat{f}(I) \neq 0$. Since $\hat{f}(I)$ is an integer multiple of $|\mathrm{rel}(f)|$, $|\hat{f}(I)| \geq 2^{-d}$. In particular, if (6) is satisfied, then

$$|\tilde{f}_S(I)| \geq |1 - 2p_I| \cdot |1 - 2\eta| \cdot |\hat{f}(I)| - \tfrac{1}{2} \cdot \gamma_a^s \cdot \gamma_b \cdot t \;\geq\; \tfrac{1}{2} \cdot \gamma_a^s \cdot \gamma_b \cdot t\,,$$

i.e., $|\beta| \geq t/2$, so x_i is classified as "relevant" with probability at least $1 - \delta/n$.

Now consider some variable $x_i \notin \mathrm{rel}(f)$. Thus $\hat{f}(I) = 0$ for all $I \subseteq [n]$ with $i \in I$ by Lemma 3.1. By (6), with probability at least $1 - \frac{\delta}{n}$, $|\tilde{f}_S(I)| \leq \frac{1}{2} \cdot \gamma_a^s \cdot \gamma_b$. We conclude that x_i is correctly classified with probability at least $1 - \delta/n$.

Finally, the probability that at least one out of the n variables is not classified correctly is at most $n \cdot (\delta/n) = \delta$. □

Note that Noisy-IRV is not only applicable to product attribute noise. The performance guaranteed by Theorem 4.9 is also valid for general distributions, provided that γ_a can be chosen such that $|1 - 2p_I| \geq \gamma_a^{|I|}$ for all $I \subseteq [n]$ with $1 \leq |I| \leq s$.

Sample complexity and running time of Noisy-IRV can be bounded as follows:

Proposition 4.10 *The algorithm* `Noisy-IRV` *has sample complexity* $O(\log(n/\delta) \cdot 2^{2d} \cdot \gamma_a^{-2s}\gamma_b^{-2})$ *and running time* $n^s \cdot \mathrm{poly}(n, 2^d, \log(1/\delta), \gamma_a^{-s}, \gamma_b^{-1})$.

4.4 Two-Phase-Learning of Juntas

The approach of learning juntas in the presence of noise is basically the same as in the noise-free case. We proceed in two phases: in the first phase, we infer all relevant variables with high probability. In the second phase, we build up the truth table of a suitable hypothesis.

The main difference to the algorithm used in the noise-free setting is that we cannot just read off the truth table from the examples since these may contain inconsistencies. Moreover, such a truth table is unlikely to be correct.

Fortunately, we have seen in Section 4.2 how to build a good hypothesis in the presence of attribute noise. The trick is that we do not apply Proposition 4.7 to the whole given sample, but restrict the sample to the variables classified as relevant in the first phase. As a consequence, the sample and time complexity for the second phase do not depend on n anymore, but only on the number d of relevant variables.

This results in an algorithm for learning the class $\mathcal{J}_d^n$ in the presence of attribute and classification noise with sample complexity growing only polynomially in $\log n$ and 2^d (instead of n^d as in Proposition 4.8). Moreover, for the subclass $\mathcal{R}_d^n(s)$, the time complexity depends on n^s instead of n^d. Precisely, the algorithm, which we call `Learn-Noisy-Juntas`, is as follows:

1. Run `Noisy-IRV`$(\delta/2, s, \gamma_a, \gamma_b)$. Let R be the set of indices of variables classified as relevant.
2. Request m examples from $EX_{P,\eta}(f)$, where $m = \mathrm{poly}(2^d, \log(2/\delta), \gamma_a^{-d}, \gamma_b^{-1})$ is the sample size as required in Proposition 4.7 with $n = d$.
3. Compute $\tilde{f}(I)$ for all $I \subseteq R$ (see (2)).
4. Output the hypothesis $h(x) = \mathrm{sgn} \sum_{I \subseteq R} \frac{\tilde{f}(I)}{(1-2p_I)\cdot(1-2\eta)} \cdot \chi_I(x)$.

Theorem 4.11 *The algorithm* `Learn-Noisy-Juntas` *exactly* (P, η)*-learns the class* $\mathcal{R}_d^n(s)$ *with confidence* $1 - \delta$ *from a sample of size* $\mathrm{poly}(\log n, 2^d, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$ *in running time* $n^s \cdot \mathrm{poly}(n, 2^d, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$.

Proof. Let $f \in \mathcal{R}_d^n(s)$. With probability at least $1 - \delta/2$, `Noisy-IRV` successfully infers the relevant variables of f. By Proposition 4.7, again with probability at least $1 - \delta/2$, hypothesis h exactly coincides with f. Hence, `Learn-Noisy-Juntas` succeeds in exactly learning the target concept with probability at least $1 - \delta$. The sample complexity can easily be derived from the above description of the algorithm. The claimed running time follows from Proposition 4.10 and Proposition 4.7. □

For the class of all d-juntas and the class of monotone d-juntas, respectively, we obtain:

Corollary 4.12 *(a) The class* $\mathcal{J}_d^n$ *can be exactly* (P, η)*-learned with confidence* $1 - \delta$ *from a sample of size* $\mathrm{poly}(\log n, 2^d, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$ *in running time* $n^d \cdot \mathrm{poly}(n, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$.

(b) The class MON_d^n *can be exactly* (P,η)*-learned with confidence* $1-\delta$ *from a sample of size* $\mathrm{poly}(\log n, 2^d, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$ *in time* $\mathrm{poly}(n, 2^d, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$.

5 Non-Uniformly Distributed Attributes

In this section we sketch how to generalize our results to product attribute distributions (not to be confused with attribute noise). We confine ourselves to presenting results for monotone functions only. The more delicate task of studying the general applicability of the methods to s-low juntas will be left for future investigations.

The examples are now distributed according to an *attribute distribution* $D : \{0,1\}^n \to [0,1]$, which we assume to be a product distribution with rates $d_1, \ldots, d_n$. Let $\sigma_i = \sqrt{d_i \cdot (1-d_i)}$ be the standard deviation of variable x_i. A learning algorithm has access to an oracle $EX_{P,\eta}(f,D)$ that first generates an attribute vector $x \sim D$ and then applies (P,η)-noise as in the uniform case. When using methods from the uniform setting, we now obtain expectations with respect to D instead of U_n. Consequently, we have to adjust the inner product on our concept space and choose an appropriate orthonormal basis, as has been proposed by Furst, Jackson, and Smith [12]. For $i \in [n]$, define $\chi_i^D : \{0,1\}^n \to \mathbb{R}$ by $\chi_i^D(x) = \frac{d_i - x_i}{\sigma_i}$. For $I \subseteq [n]$, define $\chi_I^D : \{0,1\}^n \to \mathbb{R}$ by $\chi_I^D(x) = \prod_{i \in I} \chi_i(x)$. Note that $\chi_I^{U_n} = \chi_I$. The functions $(\chi_I^D \mid I \subseteq [n])$ form an orthonormal basis with respect to the inner product $\langle f, g \rangle_D = \mathbb{E}_{x \sim D}[f(x)g(x)]$. The *D-biased Fourier coefficient of f at I* is $\hat{f}(I) = \langle f, \chi_I^D \rangle_D$, using the same notation as in the uniform case. It is not difficult to see that Lemma 3.1 generalizes to biased Fourier coefficients, paving the way to carry over techniques from the uniform setting, at least for noise-free data.

In the noisy setting, the main problem is that in general, $\chi_I^D(x \oplus a) \neq \chi_I^D(x) \cdot \chi_I^D(a)$. Hence we cannot just approximate $\mathbb{E}_{x \sim D, a \sim P, b \sim \eta}[\chi_I^D(x \oplus a) \cdot f(x) \cdot b)]$ and proceed as in the uniform case. On the other hand, using $\chi_I^{U_n}$, we obtain

$$\mathbb{E}_{x \sim D, a \sim P, b \sim \eta}[\chi_I^{U_n}(x \oplus a) \cdot f(x) \cdot b)] = (1-2p_I) \cdot (1-2\eta) \cdot \langle f, \chi_I^{U_n} \rangle_D \ ,$$

but $\langle f, \chi_I^{U_n} \rangle_D$ does not properly work together with the definition of biased Fourier coefficients. The way out is provided by a clever combination of biased Fourier coefficients, the inner product $\langle \cdot, \cdot \rangle_D$, and the "unbiased" parity functions $\chi_I^{U_n}$, presented in Lemma 5.1. Its proof relies on explicit calculations of the *biased* Fourier coefficients of the *unbiased* parity functions and the application of the identity $\langle f, \chi_I^{U_n} \rangle_D = \sum_{J \subseteq [n]} \langle f, \chi_J^D \rangle_D \langle \chi_I^{U_n}, \chi_J^D \rangle_D$.

Lemma 5.1 *Let* $f : \{0,1\}^n \to \mathbb{R}$ *and* $I \subseteq [n]$. *Then*

$$\hat{f}(I) = \left(\prod_{i \in I}(2\sigma_i)\right)^{-1} \cdot \langle f, \chi_I^{U_n} \rangle_D - \sum_{J \subsetneq I} \prod_{i \in I \setminus J} \frac{1-2d_i}{2\sigma_i} \cdot \hat{f}(J) \ .$$

Before we prove Lemma 5.1, we calculate $\langle \chi_I^{U_n}, \chi_J^D \rangle_D$. This may be of independent interest for other applications.

Lemma 5.2 *Let* $J \subseteq I \subseteq [n]$. *Then*

$$\langle \chi_I^{U_n}, \chi_J^D \rangle_D = \prod_{i \in J}(2\sigma_i) \cdot \prod_{i \in I \setminus J}(1-2d_i) \ .$$

Proof. We have

$$\chi_i^{U_n}(x) \cdot \chi_i^D(x) = (-1)^{x_i} \cdot \frac{d_i - x_i}{\sigma_i} = \begin{cases} \frac{d_i}{\sigma_i} & \text{if } x_i = 0, \\ \frac{1-d_i}{\sigma_i} & \text{if } x_i = 1. \end{cases}$$

Hence, using $\chi_I^D = \prod_{i \in I} \chi_i^D$, we obtain

$$\begin{aligned} \langle \chi_I^{U_n}, \chi_J^D \rangle_D &= \sum_{x \in \{0,1\}^n} D(x) \cdot \chi_I^{U_n}(x) \cdot \chi_J^D(x) \\ &= \sum_{x \in \{0,1\}^n} \prod_{i \in [n]: x_i = 0} (1 - d_i) \cdot \prod_{i \in [n]: x_i = 1} d_i \cdot \prod_{i \in J: x_i = 0} \frac{d_i}{\sigma_i} \cdot \prod_{i \in J: x_i = 1} \frac{1 - d_i}{\sigma_i} \cdot \prod_{i \in I \setminus J} (-1)^{x_i} \\ &= \sum_{x \in \{0,1\}^n} \prod_{i \in J} \sigma_i \cdot \prod_{i \in [n] \setminus J} (d_i^{x_i} \cdot (1 - d_i)^{1 - x_i}) \cdot \prod_{i \in I \setminus J} (-1)^{x_i} \\ &= \sum_{x \in \{0,1\}^n} \prod_{i \in J} \sigma_i \cdot \prod_{i \in [n] \setminus I} (d_i^{x_i} \cdot (1 - d_i)^{1 - x_i}) \cdot \prod_{i \in I \setminus J} ((-1)^{x_i} \cdot d_i^{x_i} \cdot (1 - d_i)^{1 - x_i}) \\ &= 2^{|J|} \cdot \prod_{i \in J} \sigma_i \cdot \sum_{x \in \{0,1\}^{I \setminus J}} \prod_{i \in I \setminus J} ((-1)^{x_i} \cdot d_i^{x_i} \cdot (1 - d_i)^{1 - x_i}) \\ &= 2^{|J|} \cdot \prod_{i \in J} \sigma_i \cdot (\mathbb{P}_{x \sim D}[\chi_{I \setminus J}^{U_n} = 1] - \mathbb{P}_{x \sim D}[\chi_{I \setminus J}^{U_n} = -1]) \\ &= 2^{|J|} \cdot \prod_{i \in J} \sigma_i \cdot (1 - 2d_{I \setminus J}) \;=\; \prod_{i \in J} (2\sigma_i) \cdot \prod_{i \in I \setminus J} (1 - 2d_i) \,, \end{aligned}$$

where analogous to p_I, we define $d_I = \mathbb{P}_{x \sim D}[\chi_I^{U_n} = -1]$ for $I \subseteq [n]$. By induction, $1 - 2d_I = \prod_{i \in I}(1 - 2d_i)$. □

Proof of Lemma 5.1. We first show that $\langle \chi_I^{U_n}, \chi_J^D \rangle_D = 0$ for all $J \not\subseteq I$:

$$\chi_I^D = \prod_{i \in I} \chi_i^D = \prod_{i \in I} (2\sigma_i)^{-1} \cdot (\chi_i^{U_n} + (2d_i - 1) \cdot 1) \in \langle \chi_J^{U_n} \mid J \subseteq I \rangle$$

implies $\langle \chi_J^D \mid J \subseteq I \rangle \subseteq \langle \chi_J^{U_n} \mid J \subseteq I \rangle$. Since both sides of this relation are subspaces of $\mathbb{R}^{\{0,1\}^n}$ of equal dimension, the spaces coincide. In particular, $\chi_I^{U_n} \in \langle \chi_J^D \mid J \subseteq I \rangle$. Consequently, $\langle \chi_I^{U_n}, \chi_J^D \rangle_D = 0$ for all $J \not\subseteq I$. Now

$$\begin{aligned} \langle f, \chi_I^{U_n} \rangle_D &= \langle f, \sum_{J \subseteq [n]} \langle \chi_I^{U_n}, \chi_J^D \rangle_D \cdot \chi_J^D \rangle_D \;=\; \sum_{J \subseteq I} \langle f, \chi_J^D \rangle_D \cdot \langle \chi_I^{U_n}, \chi_J^D \rangle_D \\ &= \sum_{J \subseteq I} \hat{f}(J) \cdot \hat{\chi}_I^{U_n}(J) \;=\; \sum_{J \subsetneq I} \hat{f}(J) \cdot \hat{\chi}_I^{U_n}(J) + \hat{f}(I) \cdot \hat{\chi}_I^{U_n}(I) \,. \end{aligned}$$

Hence

$$\hat{f}(I) = \hat{\chi}_I^{U_n}(I)^{-1} \cdot \big(\langle f, \chi_I^{U_n} \rangle_D - \sum_{J \subsetneq I} \hat{f}(J) \cdot \hat{\chi}_I^{U_n}(J) \big) \,.$$

The claim now follows from Lemma 5.2. □

The threshold to recognize nonzero Fourier coefficients is given by the least absolute value of the considered nonzero coefficients. For monotone functions, we have:

Lemma 5.3 *Let $f \in \mathcal{B}^n$ be a monotone Boolean function and $i \in [n]$. Then $x_i \in \mathrm{rel}(f)$ if and only if $\hat{f}(i) \geq 2 \cdot \prod_{x_j \in \mathrm{rel}(f)} \min\{d_j, 1-d_j\}$.*

Proof. Clearly, if $x_i \notin \mathrm{rel}(f)$, then $\hat{f}(i) = 0$, so assume that $x_i \in \mathrm{rel}(f)$. We have

$$\begin{aligned}
\hat{f}(i) &= \sum_{x \in \{0,1\}^n} D(x) \cdot f(x) \cdot \frac{d_i - x_i}{\sigma_i} \\
&= \sum_{x' \in \{0,1\}^{[n] \setminus \{i\}}} D(x') \cdot \left((1-d_i) \cdot f_{x_i=0}(x') \cdot \frac{d_i}{\sigma_i} - d_i \cdot f_{x_i=1}(x') \cdot \frac{1-d_i}{\sigma_i} \right) \\
&= \sigma_i \cdot \sum_{x' \in \{0,1\}^{[n] \setminus \{i\}}} D(x')(f_{x_i=0}(x') - f_{x_i=1}(x')) \\
&= \sigma_i \cdot \sum_{x' \in \{0,1\}^{\mathrm{rel}(f) \setminus \{i\}}} D(x')(f'_{x_i=0}(x') - f'_{x_i=1}(x')) \ ,
\end{aligned}$$

where for $x \in \{0,1\}^J$, $J \subseteq [n]$, $D(x) = \prod_{j \in J} d_i^{x_i} \cdot (1-d_i)^{1-x_i}$ and for $g : \{0,1\}^J \to \mathbb{R}$, $g' : \{0,1\}^{\mathrm{rel}(g)} \to \mathbb{R}$ denotes the restriction of g to its relevant variables. If f is monotone, then $f_{x_i=0} \geq f_{x_i=1}$ and if x_i is relevant to f, then $f_{x_i=0}(x') \neq f_{x_i=1}(x')$ for at least one $x' \in \{0,1\}^{[n] \setminus \{i\}}$. Hence

$$\hat{f}(i) \geq 2 \cdot \sigma_i \cdot \min_{x' \in \{0,1\}^{\mathrm{rel}(f) \setminus \{i\}}} D(x') = 2 \cdot \sigma_i \cdot \prod_{x_j \in \mathrm{rel}(f) \setminus \{x_i\}} \min\{d_j, 1-d_j\} \ .$$

We conclude the proof by showing $\sigma_i \geq \min\{d_i, 1-d_i\}$:
If $d_i \leq 1/2$, then $\sigma_i = \sqrt{d_i \cdot (1-d_i)} \geq d_i = \min\{d_i, 1-d_i\}$. Similarly, if $d_i \geq 1/2$, then $\sigma_i \geq 1 - d_i = \min\{d_i, 1-d_i\}$. □
The lemma also holds for locally (anti-)monotone functions with $\hat{f}(i)$ replaced by $|\hat{f}(i)|$.

Theorem 5.4 *The algorithm* `Noisy-Monotone-IRV-PDA` *shown in Fig. 2 accomplishes the following. Let $f \in \mathrm{MON}_d^n$ be a monotone d-junta. Let D be a product attribute distribution with rates $d_i \in [\gamma_c, 1-\gamma_c]$ for some $\gamma_c > 0$. Let P be a γ_a-bounded attribute noise distribution and η be a classification noise rate such that $|1-2\eta| \geq \gamma_b > 0$. Then with probability $1-\delta$, the variables classified as "relevant" by* `Noisy-Monotone-IRV-PDA` *are exactly the relevant variables of f, where the algorithm has access to $EX_{P,\eta}(f,D)$. Moreover, algorithm* `Noisy-Monotone-IRV-PDA` *has sample complexity* $\mathrm{poly}(\log n, \log(1/\delta), \gamma_a^{-1}, \gamma_b^{-1}, \gamma_c^{-d})$ *and running time* $\mathrm{poly}(n, \log(1/\delta), \gamma_a^{-1}, \gamma_b^{-1}, \gamma_c^{-d})$.

Proof. The proof is an extension of the proof of Theorem 4.9. Let S be the D-distributed (P,η)-noisy sample that the algorithm obtains from the oracle $EX_{P,\eta}(f,D)$ and let $\varepsilon = \gamma_c^d$. By Lemma 5.1,

$$\hat{f}(i) = (2\sigma_i)^{-1} \cdot \langle f, \chi_I^{U_n} \rangle_D - \frac{1-2d_i}{2\sigma_i} \cdot \hat{f}(\emptyset) \ .$$

Since $\mathbb{E}_{x \sim D, b \sim \eta}[f(x) \cdot b] = (1-2\eta) \cdot \hat{f}(\emptyset)$, it follows analogously to the proof of Lemma 4.4 that with probability at least $\delta/2n$,

$$|\phi_i - (1-2d_i) \cdot f(\emptyset)| \leq \sigma_i \cdot \varepsilon, \text{ provided that } m \geq 2 \cdot \ln\left(\frac{4n}{\delta}\right) \cdot \frac{(1-2d_i)^2}{(1-2\eta)^2 \cdot \sigma_i^2 \cdot \varepsilon^2} \ . \quad (7)$$

```
Algorithm Noisy-Monotone-IRV-PDA

input $\delta \in (0,1]$, $d_1,\ldots,d_n$, $p_1,\ldots,p_n$, $\eta$
request $m = 2\cdot\ln(\frac{4n}{\delta})\cdot\gamma_c^{-2d}\cdot(\gamma_a\cdot\gamma_b)^{-2}\cdot(\gamma_c\cdot(1-\gamma_c))^{-1}$
  examples $(x^k,y^k)_{k\in[m]}$ from $EX_{P,\eta}(f,D)$
$R \leftarrow \emptyset$
$\phi_0 \leftarrow \frac{1}{(1-2\eta)\cdot m}\sum_{k=1}^m y^k$
for $i=1$ to $n$ do
  $\phi_i \leftarrow (1-2d_i)\cdot\phi_0$
  $\psi_i \leftarrow \frac{1}{(1-2p_i)\cdot(1-2\eta)\cdot m}\sum_{k=1}^m y^k\cdot\chi_i^{U_n}(x^k)$
  $\beta_i \leftarrow \frac{\psi_i-\phi_i}{2\cdot\sqrt{d_i\cdot(1-d_i)}}$
  if $|\beta|\geq\gamma_c^d$
    then $R \leftarrow R\cup\{x_i\}$
output ``relevant variables:'' $R$
```

Figure 2: Algorithm Noisy-Monotone-IRV-PDA (IRV = Infer Relevant Variables, PDA = Product Distributed Attributes) to infer all relevant variables of monotone concepts from product distributed attributes (with rates $d_i \in [\gamma_c, 1-\gamma_c]$) in the presence of ($\gamma_a$,$\gamma_b$)-bounded noise.

Moreover, since $\mathbb{E}_{x\sim D,a\sim P,b\sim\eta}[f(x)\cdot b\cdot\chi_I^{U_n}(x\oplus a)] = (1-2p_i)\cdot(1-2\eta)\cdot\langle f,\chi_I^{U_n}\rangle_D$, with probability at least $1-\delta/2n$,

$$|\psi_i - \langle f,\chi_I^{U_n}\rangle_D| \leq \sigma_i\cdot\varepsilon, \text{ provided that } m \geq 2\cdot\ln\left(\tfrac{4n}{\delta}\right)\cdot\tfrac{1}{(1-2p_i)^2\cdot(1-2\eta)^2\cdot\sigma_i^2\cdot\varepsilon^2}\ . \quad (8)$$

Since the number of examples requested by the algorithm dominates both numbers given in (7) and (8), with probability at least $1-\delta/n$,

$$|\beta_i - \hat{f}(i)| \;=\; \left|\frac{\psi_i-\phi_i}{2\sigma_i} - \frac{\langle f,\chi_I^{U_n}\rangle_D - (1-2d_i)\hat{f}(\emptyset)}{2\sigma_i}\right| \;\leq\; \frac{2\cdot\varepsilon\cdot\sigma_i}{2\cdot\sigma_i} \;=\; \varepsilon\ .$$

Noisy-Monotone-IRV-PDA classifies x_i as "relevant" if and only if $|\beta_i|\geq\varepsilon$. If $\hat{f}(i)=0$, then $|\beta_i|\leq\varepsilon$ with probability at least $1-\delta/n$, and if $\hat{f}(i)\neq 0$, then $|\beta_i|\geq\varepsilon$ with probability at least $1-\delta/n$ (since $\hat{f}(i)\geq 2\varepsilon$ by Lemma 5.3). Consequently, all variables are classified correctly with probability at least $1-\delta$. □

Next we describe how to construct a hypothesis. We use Lemma 5.1 to successively approximate all biased Fourier coefficients level by level, i.e., given a D-distributed (P,η)-noisy sample $S=(x^k,y^k)_{k\in[m]}$ and having inferred the set R of relevant variable indices, we compute for each $I\subseteq R$ the value

$$\beta_I = \left((1-2p_I)(1-2\eta)\prod_{i\in I}2\sigma_i\right)^{-1}\cdot\frac{1}{m}\cdot\sum_{k=1}^m y^k\chi_I(x^k) - \sum_{J\subsetneq I}\prod_{i\in I\setminus J}\frac{1-2d_i}{2\sigma_i}\beta_J \quad (9)$$

and build the hypothesis $h(x)=\operatorname{sgn}\sum_{I\subseteq R}\beta_I\cdot\chi_I^D(x)$.

To ensure that β_I approximates $\hat{f}(I)$ well enough, reasonably good approximations of all coefficients $\hat{f}(J)$, $J \subseteq I$, are required. This feedback effect leads to a necessary sample size of $2^{\omega(|\operatorname{rel}(f)|)}$. In case that $|1-2d_i| \leq \sigma_i$ (which is the case if and only if $|1-2d_i| \leq 1/\sqrt{5}$, i.e., $d_i \in [0.2764, 0.7236]$), the following theorem provides upper bounds on the sample and time complexity for learning monotone juntas from product distributed examples in the presence of product attribute and classification noise:

Theorem 5.5 *Let D be a product attribute distribution with rates $d_i \in [0.2764, 0.7236]$. Let P be a γ_a-bounded product attribute noise distribution and η be a classification noise rate with $|1-2\eta| \geq \gamma_b > 0$. Then the class MON_d^n can be exactly learned under noise (P, η) with confidence $1-\delta$ from a D-distributed sample of size $\operatorname{poly}(\log n, 2^{d^2}, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$ in running time $\operatorname{poly}(n, 2^{d^2}, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$.*

Before we prove Theorem 5.5, we show that a suitable hypothesis can be build provided that the set of relevant variables is already known:

Lemma 5.6 *Let $f \in \mathcal{B}^n$ and D be a product attribute distribution such that for all $x_i \in \operatorname{rel}(f)$, $|1-2d_i| \leq 1/\sqrt{5}$. Let P be a γ_a-bounded product attribute noise distribution and η be a classification noise rate with $|1-2\eta| \geq \gamma_b > 0$. Let $R \subseteq [n]$ such that $\operatorname{rel}(f) = \{x_i \mid i \in R\}$ and let S be a D-distributed (P, η)-noisy sample of size*

$$m \geq \operatorname{poly}\left(2^{d^2}, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1}\right),$$

where $d = |\operatorname{rel}(f)|$. Let β_I as defined in (9). Then with probability at least $1-\delta$, the hypothesis h defined by

$$h(x) = \operatorname{sgn} \sum_{I \subseteq R} \beta_I \cdot \chi_I^D(x)$$

coincides with f.

Proof. We first proof by induction on $|I|$ that $|\beta_I - \hat{f}(I)| \leq \varepsilon$ provided that

$$m \geq \operatorname{poly}\left(2^{|I|^2}, \log(1/\delta), \gamma_a^{-|I|}, \gamma_b^{-1}\right). \tag{10}$$

For $|I| = 0$, $\beta_\emptyset = (1-2p_\emptyset) \cdot (1-2\eta) \cdot \frac{1}{m}\sum_{k=1}^m y^k$. By Hoeffding bounds, with probability at least $1-\delta$,

$$|\beta_\emptyset - \hat{f}(\emptyset)| \leq \varepsilon, \text{ provided that } m \geq 2 \cdot \ln\left(\frac{2}{\delta}\right) \cdot \frac{1}{(1-2\eta)^2 \cdot \varepsilon^2},$$

which is clearly dominated by (10).

Now consider $I \subseteq [n]$ with $|I| \geq 1$ and assume that the claim holds for all $J \subseteq [n]$ of size at least $|I| - 1$. Let

$$\psi_I = \left((1-2p_I) \cdot (1-2\eta) \cdot \prod_{i \in I \setminus J} 2\sigma_i\right)^{-1} \cdot \frac{1}{m}\sum_{k=1}^m y^k \cdot \chi_I^{U_n}(x^k)$$

and

$$\phi_I = \sum_{J \subsetneq I} \Big(\prod_{i \in I \setminus J} \frac{1-2d_i}{2\sigma_i} \Big) \cdot \beta_J \ .$$

Since $\mathbb{E}_{x \sim D, a \sim P, b \sim \eta}[f(x^k) \cdot b^k \cdot \chi_I^{U_n}(x^k \oplus a^k)] = (1-2p_I) \cdot (1-2\eta) \cdot \langle f, \chi_I^{U_n} \rangle_D$, with probability at least $1 - \delta \cdot 2^{-|I|}$,

$$\Big| \psi_I - \big(\prod_{i \in I} 2\sigma_i \big)^{-1} \cdot \langle f, \chi_I^{U_n} \rangle_D \Big| \leq \varepsilon/2 \ ,$$

provided that

$$m \geq 2 \cdot \ln \Big(\frac{2 \cdot 2^{|I|}}{\delta} \Big) \cdot \frac{1}{(1-2p_I)^2 \cdot (1-2\eta)^2 \cdot (\prod_{i \in I} 2\sigma_i)^2 \cdot \varepsilon^2} \ ,$$

which again is dominated by (10) since $\big(\prod_{i \in I} 2\sigma_i\big)^{-1} \in 2^{O(|I|)}$.

Furthermore, by induction hypothesis, with probability at least $1 - \delta \cdot 2^{-|I|}$,

$$|\beta_J - \hat{f}(J)| \leq \varepsilon \cdot 2^{-|J|-1} \ ,$$

provided that

$$m \geq \mathrm{poly}\Big(2^{|J|^2}, \log\Big(\frac{2^{|I|}}{\delta}\Big), \gamma_a^{-|J|}, \gamma_b^{-1} \Big) \ .$$

Therefore, since we assume that $|1 - 2d_i| \leq \sigma_i$,

$$\begin{aligned} \Big| \phi_I - \sum_{J \subsetneq I} \Big(\prod_{i \in I \setminus J} \frac{1-2d_i}{2\sigma_i} \Big) \cdot \hat{f}(J) \Big| &\leq \sum_{J \subsetneq I} \Big| \prod_{i \in I \setminus J} \frac{1-2d_i}{2\sigma_i} \Big| \cdot |\beta_J - \hat{f}(J)| \\ &\leq \sum_{J \subsetneq I} 2^{-|I \setminus J|} \cdot 2^{-|J|-1} \cdot \varepsilon = \varepsilon/2 \end{aligned}$$

with probability at least $1 - \delta \cdot \frac{2^{|I|}-1}{2^{|I|}}$, provided that

$$\begin{aligned} m &\geq \mathrm{poly}\Big(2^{(|I|-1)^2}, \log(1/\delta), |I|, \gamma_a^{-|I|}, \gamma_b^{-1} \Big)^2, \varepsilon^{-1}, 2^{|I|-1} \Big) \\ &= \mathrm{poly}\Big(2^{|I|^2}, \log(1/\delta), \gamma_a^{-|I|}, \gamma_b^{-1} \Big)^2, \varepsilon^{-1} \Big) \end{aligned}$$

for a suitable polynomial. Finally,

$$\begin{aligned} |\beta_I - \hat{f}(I)| &= \Big| \psi_I - \phi_I - \Big(\big(\prod_{i \in I} 2\sigma_i \big)^{-1} \cdot \langle f, \chi_I^{U_n} \rangle_D - \sum_{J \subsetneq I} \Big(\prod_{i \in I \setminus J} \frac{1-2d_i}{2\sigma_i} \Big) \cdot \beta_J \Big) \Big| \\ &\leq \varepsilon/2 + \varepsilon/2 \ = \ \varepsilon \end{aligned}$$

with probability at least $1 - \delta$. This finishes the induction proof.

Now we apply this result to estimate how closely h approximates f: If $|\beta_I - \hat{f}(I)| \leq \big(\prod_{i \in I} \sigma_i\big) \cdot \varepsilon / 2^d$ for all $I \subseteq R$, then

$$\begin{aligned} \Big| \sum_{I \subseteq R} \beta_I \cdot \chi_I^D(x) - \hat{f}(x) \Big| &\leq \sum_{I \subseteq R} \Big| \beta_I \cdot \hat{f}(I) \Big| \cdot \Big| \chi_I^D(x) \Big| \\ &\leq 2^{-d} \cdot \sum_{I \subseteq R} \big| \chi_I^D(x) \big| \\ &\leq 2^{-d} \cdot \sum_{I \subseteq R} \big(\prod_{i \in I} \sigma_i \big)^{-1} \ \leq \ \varepsilon \ . \end{aligned}$$

Hence, taking $\varepsilon = 1$, we have to request

$$m = \mathrm{poly}\Big(2^{d^2}, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1}, \frac{2^d}{\prod_{i\in R}\sigma_i \cdot 1}\Big)$$

examples to guarantee $h(x) = f(x)$ for all $x \in \{0,1\}^n$ with probability at least $1-\delta$. For all $i \in R$, $|1-2d_i| \le \sigma_i$ implies $\sigma_i \ge 1/\sqrt{5}$. Thus $\big(\prod_{i\in R}\sigma_i\big)^{-1} \in 2^{O(d)}$, which is absorbed by $\mathrm{poly}(2^{d^2})$. □

Now we can prove Theorem 5.5:
Proof of Theorem 5.5. By Theorem 5.4, we can infer the set of relevant attributes correctly with probability at least $1 - \delta/2$, provided that we are given a sample of size $m \ge \mathrm{poly}(\log n, \log(1/\delta), \gamma_a^{-1}, \gamma_b^{-1}, \gamma_c^{-d})$, which is dominated by the claimed sample complexity since $\gamma_c^{-d} \in 2^{O(d)}$. By Lemma 5.6, f can be exactly recovered from

$$\mathrm{poly}\Big(2^{d^2}, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1}\Big)$$

examples with probability at least $1 - \delta/2$. Combining these bounds, the claimed sample complexity follows. The claimed running time obviously suffices. □

6 Conclusion

We have investigated the learnability of Boolean juntas in the presence of attribute and classification noise. While arbitrary noise distributions may render learning impossible, we have presented an algorithm to learn the class of s-low d-juntas under product attribute and classification noise with rates different from 1/2. For $s = 1$, these include all monotone juntas. Moreover, the algorithm does not only work for product noise distributions but for any distribution satisfying a more general condition (as stated in (4)). In addition, we have shown how to generalize the methods to non-uniformly distributed examples.

The major goal is to settle the question whether learning juntas in the presence of noise can be done as efficiently (up to unavoidable factors due to noise) as in the noise-free case. At present, this means whether or not running time $n^{c\cdot d} \cdot \mathrm{poly}(n, 2^d, \gamma_a^d, \gamma_b^{-1})$ can be achieved for learning $\mathcal{J}_d^n$, with some constant $c < 1$ ($c < 0.704$ would even improve the noise-free case). While we have shown that the "Fourier part" of Mossel et al. [22] carries over to the noisy scenario, it seems that an adaption of the "parity part" is intractable since it requires noise-tolerant learning of parity functions. We suspect that non-trivial lower bounds (based on hardness assumptions) can be shown.

References

[1] T. Akutsu, S. Miyano, and S. Kuhara. Algorithms for Identifying Boolean Networks and Related Biological Networks Based on Matrix Multiplication and Fingerprint Function. *J. Comput. Biology*, 7(3-4):331–343, 2000.

[2] D. Angluin and P. Laird. Learning From Noisy Examples. *Machine Learning*, 2(4):343–370, 1988.

[3] J. Arpe and R. Reischuk. Robust Inference of Relevant Attributes. In Proc. ALT 2003, LNCS 2842, 99–113.

[4] J. Arpe and R. Reischuk. Robust Inference of Relevant Attributes. Tech. Rep. SIIM-A-03-12, Universität zu Lübeck, 2003.

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

[6] A. Bernasconi. *Mathematical Techniques for the Analysis of Boolean Functions*. PhD thesis, Università degli Studi di Pisa, Dipartimento di Ricerca in Informatica, 1998.

[7] A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271, 1997.

[8] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Occam's razor. *Inform. Process. Lett.*, 24(6):377–380, 1987.

[9] N. Bshouty, J. Jackson, and C. Tamon. Uniform-distribution attribute noise learnability. *Information and Computation*, 187(2):277–290, 2003.

[10] S. Decatur and R. Gennaro. On Learning from Noisy and Incomplete Examples. In Proc. COLT 1995, 353–360.

[11] D. Fukagawa and T. Akutsu. Performance Analysis of a Greedy Algorithm for Inferring Boolean Functions. *Inform. Process. Lett.*, 93(1):7–12, January 2005.

[12] Merrick L. Furst, Jeffrey C. Jackson, and Sean W. Smith. Improved Learning of AC^0 Functions. In Proc. COLT 1991, 317–325.

[13] S. Goldman and R. Sloan. Can PAC learning algorithms tolerate random attribute noise? *Algorithmica*, 14(1):70–84, 1995.

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

[15] J. Kahn, G. Kalai, and N. Linial. The Influence of Variables on Boolean Functions (extended abstract). In Proc. FOCS 1988, 68–80.

[16] M. Kolountzakis, E. Markakis, and A. Mehta. Learning symmetric k-juntas in time $n^{o(k)}$. arXiv:math.CO/0504246 v1, 2005. `http://arxiv.org/abs/math.CO/0504246v1`.

[17] N. Linial, Y. Mansour, and N. Nisan. Constant Depth Circuits, Fourier Transform, and Learnability. *J. ACM*, 40(3):607–620, 1993.

[18] R. Lipton, E. Markakis, A. Mehta, and N. Vishnoi. On the Fourier Spectrum of Symmetric Boolean Functions with Applications to Learning Symmetric Juntas. To appear in Proc. CCC 2005.

[19] N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1987.

[20] A. Miyata, J. Tarui, and E. Tomita. Learning Boolean Functions in AC^0 on Attribute and Classification Noise. In Proc. ALT 2004, LNAI 3244, 142–155.

[21] E. Mossel and R. O'Donnell. On the Noise Sensitivity of Monotone Functions. *Random Structures Algorithms*, 23(3):333–350, 2003.

[22] E. Mossel, R. O'Donnell, and R. Servedio. Learning functions of k relevant variables. *J. Comput. System Sci.*, 69(3):421–434, 2004.

[23] R. O'Donnell. *Computational Applications Of Noise Sensitivity.* PhD thesis, Department of Mathematics, Mass. Inst. Tech., 2003.

[24] G. Shackelford and D. Volper. Learning k-DNF with Noise in the Attributes. In Proc. COLT 1988, 97–103.