Available online at www.sciencedirect.com
ScienceDirect
Theoretical Computer Science 384 (2007) 2–21
Theoretical Computer Science
www.elsevier.com/locate/tcs

Learning juntas in the presence of noise☆

Jan Arpe*, Rüdiger Reischuk

Institut für Theoretische Informatik, Universität zu Lübeck, Ratzeburger Allee 160, 23538 Lübeck, Germany

Abstract

We investigate the combination of two major challenges in computational learning: dealing with huge amounts of irrelevant information and learning from noisy data. It is shown that large classes of Boolean concepts that depend only on a small fraction of their variables – so-called *juntas* – can be learned efficiently from uniformly distributed examples that are corrupted by random attribute and classification noise. We present solutions to cope with the manifold problems that inhibit a straightforward generalization of the noise-free case. Additionally, we extend our methods to non-uniformly distributed examples and derive new results for monotone juntas and for parity juntas in this setting. It is assumed that the attribute noise is generated by a product distribution. Without any restrictions of the attribute noise distribution, learning in the presence of noise is in general impossible. This follows from our construction of 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; Learning in the presence of noise; Learning in the presence of irrelevant information; Juntas; Fourier analysis

1. Introduction

1.1. Motivation

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 can be done if both phenomena occur: How can we learn n-ary Boolean concepts that depend on only a small number d of unknown 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 [8]). 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

☆ This research was supported by DFG research grant Re 672/4.
* Corresponding author. Tel.: +49 451 500 5312; fax: +49 451 500 5301.
E-mail addresses: arpe@tcs.uni-luebeck.de (J. Arpe), reischuk@tcs.uni-luebeck.de (R. Reischuk).

0304-3975/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2007.05.014

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 $n^{0.704d} \cdot \text{poly}(n, 2^d, \log(1/\delta))$, 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). The Fourier method yields an algorithm for learning the class of monotone d-juntas in time $\text{poly}(n, 2^d, \log(1/\delta))$. Learning juntas is also closely related to other highly important open questions in learning theory: learning $\omega(1)$-sized decision trees or DNF formulas in polynomial time is equivalent to learning $\omega(1)$-juntas in polynomial time; see Mossel et al. [22] for more details on this issue. While learning arbitrary k-term DNFs in polynomial time might be a too hard goal to achieve, there are positive results for learning *monotone* juntas. This may indicate that efficiently learning *monotone* DNF in polynomial time might indeed be possible. See Servedio [24] for a survey on results concerning the latter problem.

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 [3] 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 [25] and later by Decatur and Gennaro [12]. Bshouty, Jackson, and Tamon [11] 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.

1.2. Our contribution

Our main contribution is a method 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 τ-*low* d-*juntas* is exactly learnable from $\text{poly}(\log n, 2^d, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$ examples in time $n^\tau \cdot \text{poly}(n, 2^d, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$ under (γ_a, γ_b)-bounded noise. Roughly speaking, a concept is τ-low if it suffices to check all Fourier coefficients up to the τth level in order to find all relevant attributes. As a main application, the class of monotone d-juntas, for which $\tau = 1$, is learnable in time $\text{poly}(n, 2^d, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$ under (γ_a, γ_b)-bounded noise.

How much do our algorithms have to know about the noise distributions? To infer the relevant attributes, lower bounds on γ_a, γ_b suffice. In order to additionally output a matching hypothesis, the attribute noise distribution has to be known exactly (or at least approximated reasonably well, see [11]). For the classification noise parameter γ_b, it still suffices to have some lower bound. Miyata et al. [20] showed how to learn the class $\mathcal{AC}^0$ in quasipolynomial time under product attribute and classification noise with $p_1 = \cdots = p_n = \eta$ without any prior knowledge of $\eta < 1/2$. On the other hand, Goldman and Sloan [14] proved that under unknown product attribute noise, learning any non-trivial concept class with accuracy ε (which is 2^{-d} for exactly learning d-juntas) is only possible if $p_i < 2\varepsilon$ for all i. If the noise distribution can be arbitrary and is *unknown* to the learner, then learning non-trivial classes is impossible; see [11].

1.3. Our techniques

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 $n^d \cdot \text{poly}(n, 2^d, \log(1/\delta))$ 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 simply read off a truth table from the undisturbed examples. This is impossible in the 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 [18] and extended to the noisy scenario by Bshouty, Jackson, and Tamon [11]. 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 $m = \mathrm{poly}(\log n, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1}, \gamma_c^{-d^2})$ examples in time $\mathrm{poly}(m, n)$, and parity d-juntas are learnable from $m = \mathrm{poly}(\log n, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1}, \gamma_c^{-1}, \theta^{-d})$ examples in time $\mathrm{poly}(m, n)$, provided that $|1-2d_i| \geq \theta > 0$ for all $i \in \{1, \ldots, n\}$. 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 longer 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 $\hat{f}(I)$, $I \subseteq \{1, \ldots, n\}$, 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 and parity juntas.

Concerning the probabilities d_i, we assume that these are exactly known to the learner, even though close approximations to these rates would certainly suffice (but would make the analysis even more technical). Such approximations could be obtained by sampling (unlabeled) examples from a noise-free source (or even from a noisy source, concluding the noise-free rates by a short calculation that involves the known noise parameters).

Finally, we prove that without restricting the attribute noise distributions (for example to product distributions), noise-tolerant learning is in general impossible, even if the noise distribution is completely known: 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 cannot be extended to arbitrary noise distributions.

Our proofs have three main ingredients: standard Hoeffding bounds [15], harmonic analysis of Boolean functions under uniform [7] and non-uniform [4,13,24] distribution, and a *noise operator*. The latter is a generalization of the *Bonami–Beckner operator*, which plays an important role in various contexts [10,5,16,6].

1.4. Organization of this paper

In Section 2, we introduce basic notation, definitions, and tools. The learning and noise models under consideration are introduced in Section 3. After reviewing how to learn juntas in the noise-free case in Section 4, we provide in Section 5 the main tools used to derive results in the noisy scenario: the noise operator and the approximation of Fourier coefficients from noisy examples. In addition, that section contains two upper learning bounds and an impossibility result. In Section 6, we show how to learn the relevant variables in the noisy case. The construction of a suitable hypothesis from noisy examples is described in Section 7. Section 8 deals with the extension of the model and tools to non-uniformly distributed attributes. In Section 9, we show how to learn the relevant variables from non-uniformly distributed noisy samples. Subsequently, in Section 10, we show how to construct a hypothesis under these conditions.

2. Preliminaries

We consider Boolean functions $f : \{0,1\}^n \to \{-1,+1\}$, also called *concepts*. The variables of f are also referred to as *attributes*. A concept is *monotone* if for all $x, y \in \{0,1\}^n$ such that $x \leq y$, we have $f(x) \geq f(y)$ (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 $I \subseteq [n] = \{1, \ldots n\}$, we define the *parity function* $\chi_I : \{0,1\}^n \to \{-1,+1\}$

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 componentwise *exclusive or*. We denote probabilities by Pr and expectations by $\mathbb{E}$. The uniform distribution on $\{0,1\}^n$ is denoted by U_n, i.e., $U_n(x) = 2^{-n}$ for all $x \in \{0,1\}^n$. We indicate by $X \sim D$ that the random variable X is distributed according to the distribution D. Moreover, if X only takes values -1 and $+1$ with $\Pr[X = -1] = p$, we write $X \sim p$. The sign function $\operatorname{sgn} : \mathbb{R} \to \{-1,+1\}$ is defined by $\operatorname{sgn}(x) = -1$ if $x < 0$ and $\operatorname{sgn}(x) = +1$ if $x \geq 0$. In particular, we define $\operatorname{sgn}(0) = +1$ for technical reasons. The functions log and ln denote the binary and the natural logarithm, respectively.

A *concept class* is a set of concepts $f : \{0,1\}^n \to \{-1,+1\}$. Let $\mathcal{C}$ be a concept class and $f \in \mathcal{C}$. A vector $(x_1, \ldots, x_n, y) \in \{0,1\}^n \times \{-1,+1\}$ is called an *example*. It is *consistent with* f if $f(x_1, \ldots, x_n) = y$. 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 [7].

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

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

If $I = \{i\}$, we write $\hat{f}(i)$ instead of $\hat{f}(\{i\})$. Intensively used features of Fourier analysis are the *Fourier expansion*

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

for all $x \in \{0,1\}^n$ and *Parseval's equation*

$$\sum_{I\subseteq[n]} \hat{f}(I)^2 = \|f\|_2^2 = 2^{-n} \sum_{x\in\{0,1\}^n} f(x)^2. \tag{2}$$

The following is a well-known technical tool to bound the probability of deviations of the statistical mean from the expected value in sufficiently large samples; see also Alon and Spencer [2]:

Lemma 2.1 (*Hoeffding Bound [15]*). *Let X_i, $i \in [n]$, be mutually independent random variables taking values in the real interval $[a, b]$, $a < b$. Then for any $\varepsilon \in [0,1]$,*

$$\Pr\left[\left|\sum_{i=1}^n X_i - \sum_{i=1}^n \mathbb{E}[X_i]\right| \geq \varepsilon n\right] \leq 2\exp\left(\frac{-2n\varepsilon^2}{(b-a)^2}\right).$$

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

$$\tilde{f}_S(I) = \frac{1}{m} \sum_{k=1}^m \chi_I(x^k) \cdot y^k. \tag{3}$$

By Lemma 2.1, if $y^k = f(x^k)$ for all $k \in [m]$, then $\tilde{f}_S(I)$ 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 $f : \{0,1\}^n \to \{-1,+1\}$ *depends* on variable x_i (and x_i is *relevant* to f) if the $(n-1)$-ary subfunctions $f_{x_i=0}$ and $f_{x_i=1}$ with x_i set to 0 and 1, respectively, are not equal. Equivalently, x_i is relevant to f if and only if there exists an $x \in \{0,1\}^n$ such that $f(x) \neq f(x \oplus e_i)$, where $e_i \in \{0,1\}^n$ denotes the vector with a one at the ith position and zeros at all other positions. Denote the set of relevant variables of f by $\operatorname{rel}(f)$. A function that depends on at most d variables is called a *d-junta*, and the class of n-ary 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. A parity function χ_I with $|I| \leq d$ is called a *parity d-junta*. The class of parity d-juntas defined on n variables is denoted by PAR_d^n.

3. Learning and noise models

Fix a *target concept* $f : \{0,1\}^n \to \{-1,+1\}$, an *attribute noise distribution* $P : \{0,1\}^n \to [0,1]$, and a *classification noise rate* $\eta \in [0,1]$.

Definition 3.1 ((P, η)-*Noisy Sample*). Let $x \sim D$, $a \sim P$, and $b \in \{-1, +1\}$ with $b \sim \eta$. The pair $(x \oplus a, f(x) \cdot b)$ is called a D-*distributed* (P, η)-*noisy example* for f. A sequence S of m independent D-distributed (P, η)-noisy examples for f is called a D-*distributed* (P, η)-*noisy sample* for f of size m.

In other words, a (P, η)-noisy example is obtained from a noise-free example (x, y) by adding a noise-vector $a \sim P$ to the attribute vector x (componentwise modulo 2) and flipping the classification y according to the classification noise bit $b \sim \eta$.

A $(P, 0)$-*noisy example* is corrupted only by attribute noise but not by classification noise. Note that noise-free examples are a special case of noisy examples: choose $P(0^n) = 1$ and $P(x) = 0$ for $x \neq 0^n$ and $\eta = 0$.

Definition 3.2 (*Learning Algorithm*). Let $\delta \in (0, 1]$ and $\varepsilon \in [0, 1]$, called the *confidence parameter* and the *accuracy parameter*, respectively. An algorithm $\mathcal{A}$ *learns* the class $\mathcal{C}$ with confidence $1-\delta$ and accuracy $1-\varepsilon$ from D-distributed (P, η)-noisy samples of size m if the following is satisfied. For all target concepts $f \in \mathcal{C}$, given a D-distributed (P, η)-noisy sample S of size m as input, $\mathcal{A}$ outputs a concept $h : \{0, 1\}^n \to \{-1, +1\}$ such that with probability at least $1 - \delta$ (taken over the set of D-distributed (P, η)-noisy samples of size m), h is ε-close to f, i.e.,

$$\Pr_{x \sim D}[h(x) \neq f(x)] \leq \varepsilon.$$

The concept h is called the *hypothesis* of $\mathcal{A}$ on input S. Algorithm $\mathcal{A}$ is a *distribution-free* learning algorithm if it learns $\mathcal{C}$ for arbitrary attribute distributions D, without any a priori knowledge about D. This is the original definition of *PAC learnability* introduced by Valiant [26]. Learning with accuracy $\varepsilon = 0$ is referred to as *exact learning*. The sample size m needed by $\mathcal{A}$ to learn (with a certain confidence and a certain accuracy) is called the *sample complexity* of $\mathcal{A}$. It is a function of the parameters δ, ε, P, η, $\mathcal{C}$, and n.

Definition 3.3 (*Learnability of a Concept Class*). A concept class $\mathcal{C}$ is *learnable* (with confidence $1-\delta$ and accuracy $1 - \varepsilon$ from D-distributed (P, η)-noisy samples of size m in time t) if there exists an algorithm $\mathcal{A}$ that learns $\mathcal{C}$ (with confidence $1 - \delta$ and accuracy ε from D-distributed (P, η)-noisy samples of size m in time t). It is *exactly learnable* if it is learnable with accuracy 1.

For the time being, we restrict ourselves to uniformly distributed attribute values. The case of non-uniform distributions is discussed in Sections 8–10.

Since arbitrary attribute noise distributions often turn out to make learning impossible, we also study the more restricted *product random attribute noise* considered by Goldman and Sloan [14]. 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

$$P(a_1, \ldots, a_n) = \prod_{i:a_i=1} p_i \cdot \prod_{i:a_i=0} (1 - p_i) = \prod_{i=1}^{n} p_i^{a_i} \cdot (1 - p_i)^{1-a_i}.$$

Naturally, such product distributions P induce product distributions on the subcubes $\{0, 1\}^I$, $I \subseteq [n]$, which we denote by P again. In general, given a product distribution P on $\{0, 1\}^n$, we refer to the probabilities $p_i = \Pr[x_i = 1]$ as the *rates* of P. In many situation, it is desirable to bound the rates away from $1/2$:

Definition 3.4 (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 $i \in [n]$, $|1 - 2p_i| \geq \gamma_a$.

If $\eta = 1/2$, then the corrupted classifications are purely random and thus not at all correlated with f. Hence, in this situation, learning is impossible. Consequently, we assume that there exists some bound $\gamma_b > 0$ such that $|1 - 2\eta| \geq \gamma_b$.

4. Review of the noise-free case

In this section, we review the "Fourier algorithm" for the noise-free scenario, as 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 τ-*low juntas* and to understand the methods presented in Section 5.

Algorithm 1 τ-FOURIER$_d$

```
input S = ((x_1^k, ..., x_n^k), y^k)_{k∈[m]}
R ← ∅
for I ⊆ [n] with 1 ≤ |I| ≤ τ do
   β ← (1/m) · Σ_{k=1}^{m} χ_I(x^k) · y^k
   if |β| ≥ 2^{-d-1}
     then R ← R ∪ {x_i | i ∈ I}
output τ-FOURIER_d(S) = R
```

Let $f : \{0,1\}^n \to \{-1,+1\}$ 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 $\hat{f}(i) = \Pr_{x\sim U_n}[f(x) = x_i] - \Pr_{x\sim U_n}[f(x) \neq x_i] \neq 0$. 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 (3). 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. [9]), there is only one hypothesis having the same set of relevant variables and being consistent with the examples, 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 lemma (implicitly used in Mossel et al. [22]) to extend the method to larger classes of Boolean concepts by looking beyond the first level of Fourier coefficients. Intuitively, the lemma says that a variable x_i is relevant to a concept f if and only if f has nonzero correlation with at least one of the parity functions χ_I with $i \in I$.

Lemma 4.1. *Let $f : \{0,1\}^n \to \{-1,+1\}$. 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 $d = |\mathrm{rel}(f)|$. However, there are $\Theta(n^d)$ such subsets, an amount that we would like to reduce. This leads us to:

Definition 4.2 (τ-*lowness*). Let $f \in \mathcal{J}_d^n$, $x_i \in \mathrm{rel}(f)$, and $\tau \in [d]$. Variable x_i is τ-*low for* f if there exists an $I \subseteq [n]$ such that $i \in I$, $|I| \leq \tau$, and $\hat{f}(I) \neq 0$. The concept f is τ-*low* if all $x_i \in \mathrm{rel}(f)$ are τ-low for f. The set of τ-low d-juntas is denoted by $\mathcal{J}_d^n(\tau)$.

In these terms, monotone juntas are 1-low, i.e., $\mathrm{MON}_d^n \subseteq \mathcal{J}_d^n(1)$. Even more: all unate juntas 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 $\mathcal{J}_d^n(1)$ since a random junta fulfills $\hat{f}(i) \neq 0$ for all $x_i \in \mathrm{rel}(f)$ with overwhelming probability; see Blum and Langley [8] and Mossel et al. [22].

Also for other subclasses $\mathcal{C}$ of $\mathcal{J}_d^n$, finding the smallest τ such that $\mathcal{C} \subseteq \mathcal{J}_d^n(\tau)$ has recently attracted considerable interest. The class of all unbalanced d-juntas is contained in $\mathcal{J}_d^n((2/3)\cdot d)$ (see Mossel et al. [22]), and the class $\mathrm{SYM}_d^n \setminus \mathrm{PAR}_d^n$ of symmetric d-juntas that are not parity functions is now known to be contained in $\mathcal{J}_d^n(O(d/\log d))$ (see Kolountzakis et al. [17]).

The algorithm for inferring the relevant variables of τ-low d-juntas (which we call τ-FOURIER$_d$) described by Mossel et al. [22] is presented as Algorithm 1.

Proposition 4.3 (*[22]*). *Let $f \in \mathcal{J}_d^n(\tau)$ be a τ-low d-junta. Then on input S, τ-FOURIER$_d(S)$ outputs exactly the relevant variables of f from a sample of size* $\mathrm{poly}(\log n, 2^d, \log(1/\delta))$ *in time* $n^\tau \cdot \mathrm{poly}(n, 2^d, \log(1/\delta))$ *with probability at least* $1-\delta$.

5. Tools for the noisy case: Uniformly distributed attributes

Now let us see what we can do if the examples contain errors. Throughout the remainder of this section, we fix an attribute noise distribution $P : \{0,1\}^n \to [0,1]$ and a classification noise rate $\eta \in [0,1]$.

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.,

$$p_I = \Pr_{a\sim P}[\chi_I(a) = -1], \tag{4}$$

and let $\lambda_I = \mathbb{E}_{a\sim P}[\chi_I(a)] = 1 - 2p_I$.

Furthermore, for the rest of this section, we fix a confidence parameter $\delta \in (0,1]$, an accuracy parameter $\varepsilon \in (0,1]$, and a target concept $f : \{0,1\}^n \to \{-1,+1\}$. Let S denote a uniformly distributed (P,η)-noisy sample of size m for f. All probabilities are taken over the possible outcomes of S for a fixed sample size m.

5.1. The noise operator

We now introduce a mathematical tool that will be used to prove upper and lower sample bounds:

Definition 5.1 (*Noise Operator*). Let $P : \{0,1\}^n \to [0,1]$ be an attribute noise distribution. We define the *noise operator* $T_P : \mathbb{R}^{\{0,1\}^n} \to \mathbb{R}^{\{0,1\}^n}$ by

$$T_P(f)(x) = \mathbb{E}_{a\sim P}[f(x \oplus a)] \tag{5}$$

for $f : \{0,1\}^n \to \mathbb{R}$ and $x \in \{0,1\}^n$.

For $f : \{0,1\}^n \to \{-1,+1\}$, $T_P(f)(x)$ may be interpreted as follows. If x is a noise-free attribute vector that is drawn according to uniform distribution, then $T_P(f)(x)$ is the expected value of the classification of the corrupted attribute vector $x \oplus a$. The function $T_P(f)$ may be thought of as the bias of a probabilistic concept: on input $x \in \{0,1\}^n$, the outcome is -1 with probability $(1 - T_P(f)(x))/2$ and $+1$ with probability $(1 + T_P(f)(x))/2$. Learning from noisy examples thus means to learn the target concept f, even though only examples of this probabilistic concept are available. By linearity of expectation, T_P is a linear operator.

For the special case that P is a product distribution with rates $p_1 = \cdots = p_n$, this operator has been extensively studied in the literature, e.g., by Kahn, Kalai, and Linial [16], Benjamini et al. [6], Mossel and O'Donnell [21], and O'Donnell [23].

We show how the Fourier coefficients of $T_P(f)$ are related to those of f.

Lemma 5.2. *Let $f : \{0,1\}^n \to \mathbb{R}$, P be an attribute noise distribution, and $I \subseteq [n]$. Then*

(a) $T_P(\chi_I) = \lambda_I \chi_I$ *and*
(b) $\widehat{T_P(f)}(I) = \lambda_I \hat{f}(I)$.

Proof. (a) For all $x \in \{0,1\}^n$, we have

$$T_P(\chi_I)(x) = \mathbb{E}_{a\sim P}[\chi_I(x \oplus a)] = \mathbb{E}_{a\sim P}[\chi_I(x) \cdot \chi_I(a)] = \lambda_I \cdot \chi_I(x).$$

(b) By linearity of the Fourier transform and T_P, we have

$$\widehat{T_P(f)}(I) = \sum_{J\subseteq[n]} \hat{f}(J)\widehat{T_P(\chi_J)}(I) = \sum_{J\subseteq[n]} \hat{f}(J)\lambda_J\widehat{\chi_J}(I) = \lambda_I \hat{f}(I). \quad \square$$

Using the Fourier expansion (1) and Parseval's equality (2), the following corollary is immediate:

Corollary 5.3. *Let $f : \{0,1\}^n \to [-1,1]$ and $P : \{0,1\}^n \to [0,1]$ be an attribute noise distribution. Then*

(a) $\|T_P(f)\|_2^2 = \mathbb{E}_{x\sim U_n}[(\mathbb{E}_{a\sim P}[f(x \oplus a)])^2] = \sum_{I\subseteq[n]} \lambda_I^2 \hat{f}(I)^2$.
(b) $\|T_P(f)\|_2^2 \le \|T_P(f)\|_1 \le \|T_P(f)\|_2$.
(c) $\|T_P(f)\|_2^2 \ge \min_{I\subseteq[n]} \lambda_I^2 \cdot \|f\|_2^2$.

Proof. Part (a) follows from Lemma 5.2(b) and from Parseval's equality (2). The first inequality of part (b) follows since, for all $g : \{0,1\}^n \to [-1,+1]$, we have

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

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 (b) follows from $\mathbb{E}[|X|]^2 \le \mathbb{E}[X^2]$ for real-valued random variables X. Finally, part (c) is an immediate consequence of part (a). □

For any $\varepsilon > 0$, Bshouty et al. [11] have defined $\Delta_P^\varepsilon(\mathcal{C})$ to be the minimum noisy distance between ε-far concepts inside $\mathcal{C}$. In terms of the noise operator, this is

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

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

One of their main results [11, Theorem 3], which is also used in our proofs, easily follows from Corollary 5.3:

Theorem 5.4 (*[11]*). *Let $P : \{0,1\}^n \to [0,1]$ be a probability distribution and $f,g : \{0,1\}^n \to \{-1,+1\}$. Then $\frac{1}{2}\|T_P(f-g)\|_2^2 \le \|T_P(f-g)\|_1 \le \|T_P(f-g)\|_2$.*

5.2. Approximating Fourier coefficients from noisy samples

Given a uniformly distributed (P,η)-noisy sample, the empirical Fourier coefficient $\tilde{f}_S(I)$ approximates

$$\mathbb{E}_{x\sim U_n, a\sim P, b\sim\eta}[\chi_I(x\oplus a)\cdot f(x)\cdot b]. \tag{6}$$

Since $\chi_I(x\oplus a) = \chi_I(x)\cdot\chi_I(a)$ and since x, a, and b are assumed to be independent, the expectation (6) equals

$$\mathbb{E}_{a\sim P}[\chi_I(a)]\cdot\mathbb{E}_{b\sim\eta}[b]\cdot\mathbb{E}_{x\sim U_n}[f(x)\cdot\chi_I(x)] \;=\; (1-2p_I)\cdot(1-2\eta)\cdot\hat{f}(I),$$

with p_I as defined in (4) (this calculation has also been carried out by Bshouty et al. [11, Proof of Theorem 8]). Using the Hoeffding bound (Lemma 2.1), we obtain

Lemma 5.5. *Let $m \ge 2\cdot\ln(2/\delta)\cdot(1/\varepsilon^2)$. 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 $\hat{f}(I)$ by dividing $\tilde{f}_S(I)$ by $(1-2p_I)(1-2\eta)$. This is possible if and only if $p_I \ne 1/2$ and $\eta \ne 1/2$. Requesting η to be different from $1/2$ is reasonable, as we have discussed above. Unfortunately, it can happen that $p_I = 1/2$ for some I (even if $\Pr_{a\sim P}[a_i = -1] \ne 1/2$ for all $i \in [n]$), yielding a concept class $\mathcal{C}$ and an attribute noise distribution P such that $\mathcal{C}$ is (information-theoretically) not $(P,0)$-learnable:

Theorem 5.6. *There is a concept class $\mathcal{C}$ and an attribute noise distribution P such that $\mathcal{C}$ is not $(P,0)$-learnable. In addition, P may be chosen such that $p_{\{i\}} < 1/2$ for all $i \in [n]$.*

Proof. Let $n = 2$ and $P : \{0,1\}^2 \to [0,1]$ be defined by

$$P(00) = 1/2,\;\; P(01) = 1/4,\;\; P(10) = 1/4,\;\text{ and }\; P(11) = 0.$$

Then $p_{\{1\}} = P(10) + P(11) = 1/4$ and $p_{\{2\}} = P(01) + P(11) = 1/4$. Let $f(x) = \chi_{\{1,2\}}(x) = (-1)^{x_1+x_2}$ and $\mathcal{C} = \{f, -f\}$. Then for each $x \in \{0,1\}^2$,

$$\Pr_{a\sim P}[f(x\oplus a) = -1] = 1/2 = \Pr_{a\sim P}[-f(x\oplus a) = -1],$$

and $f(x\oplus a)$ is independent of x. 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. This implies that $(x\oplus a, f(x))$ and $(x\oplus a, -f(x))$ are also identically distributed since $(x\oplus a, f(x)) \sim (x, f(x\oplus a))$. Hence, f and $-f$ are information-theoretically indistinguishable under P-attribute noise. □

The proof of the previous theorem demonstrates that it may happen that the parity of x_1 and x_2 changes with probability $1/2$, although each attribute separately is flipped with probability strictly less than $1/2$. In this case, the uncorrupted value of the parity $x_1 \oplus x_2$ is no longer recoverable from any number of P-noisy attribute vectors.

In contrast, things look much nicer for product distributions P with noise rates p_i that are all different from $1/2$. It is easy to prove by induction that γ_a-bounded product distributions satisfy

$$\forall I \subseteq [n] : |1 - 2p_I| \geq \gamma_a^{|I|}. \tag{7}$$

From now on, we restrict ourselves to γ_a-bounded product distributions. However, all results extend to arbitrary distributions for which condition (7) holds.

If all $p_I \neq 1/2$ and $\eta \neq 1/2$, 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

$$h(x) = \operatorname{sgn} \sum_{I \subseteq [n]} \frac{\tilde{f}(I)}{(1 - 2p_I) \cdot (1 - 2\eta)} \cdot \chi_I(x). \tag{8}$$

Precisely, Bshouty et al. [11, Theorem 8] have shown:

Proposition 5.7 *([11]). Let $\mathcal{C}$ be a concept class that is closed under complement (in the sense that $f \in \mathcal{C}$ implies $-f \in \mathcal{C}$) and $\varepsilon > 0$ such that there exists a set $\mathcal{T}_\varepsilon \subseteq 2^{[n]}$ with $\sum_{I \in \mathcal{T}_\varepsilon} \hat{f}(I)^2 \geq 1 - \varepsilon$ for all $f \in \mathcal{C}$ and $\{\chi_I \mid I \in \mathcal{T}_\varepsilon\} \subseteq \mathcal{C}$. Then for every $\delta > 0$, $\mathcal{C}$ is learnable with confidence $1 - \delta$ and accuracy $1 - 2\varepsilon$ from uniformly distributed (P, η)-noisy samples in time polynomial in $|\mathcal{T}_\varepsilon|$, $1/\Delta_P^\varepsilon(\mathcal{C})$, $\log(1/\delta)$, $1/\varepsilon$, and $1/|1 - 2\eta|$.*

For learning the class of *all* n-ary concepts, we obtain a sample and a time complexity as follows:

Proposition 5.8. *Let $\mathcal{C}$ be the class of all n-ary concepts, 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* $\operatorname{poly}(2^n, \log(1/\delta), \gamma_a^{-n}, \gamma_b^{-1})$.

Proof. By Proposition 5.7, choosing $\varepsilon = 2^{-n-1}$ and $\mathcal{T}_\varepsilon = 2^{[n]}$, 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 5.2 (see (7)), $|\lambda_I| = |1 - 2p_I| \geq \gamma_a^{|I|}$.

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 Corollary 5.3 (c), we have

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

By Theorem 5.4, $\frac{1}{2}\|T_P(f - g)\|_1 \geq 2^{-n}\gamma_a^{2n}$, 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 Proposition 5.7. □

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 7).

Since d-juntas have all of their Fourier weight located in levels $0, \ldots, d$ (by Lemma 4.1), 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 Eq. (8).

Proposition 5.9. *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* $n^d \cdot \operatorname{poly}(n, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$.

Proof. We proceed similarly as in the proof of Proposition 5.8, but choose $\varepsilon = 2^{-d-1}$ and $\mathcal{T}_\varepsilon = \{I \subseteq [n] \mid |I| \leq d\}$ (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 5.8) from below. By (7), $|\lambda_I| = |1 - 2p_I| \geq \gamma_a^{|I|}$. 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

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

Algorithm 2 τ-NOISY-FOURIER$_d$.

```
1: input S = ((x_1^k, ..., x_n^k), y^k)_{k∈[m]}, γ_a, γ_b
2: R ← ∅
3: for I ⊆ [n] with 1 ≤ |I| ≤ τ do
4:    β ← (γ_a^{|I|} · γ_b)^{-1} · (1/m) · Σ_{k=1}^m χ_I(x^k) · y^k
5:    if |β| ≥ 2^{-d-1}
6:      then R ← R ∪ {x_i | i ∈ I}
7: output τ-NOISY-FOURIER_d(S) = R
```

By Theorem 5.4, $\frac{1}{2}\|f-g\|_1 \geq 2^{-d-1}\cdot\gamma_a^{4d}$, 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 Proposition 5.7. □

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 is the case for example for monomials; see e.g. [23, Section 3.3]).

In what follows 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 7.1, we show that this significantly reduces the sample complexity from $O(n^{d+O(1)})$ to poly$(\log n, 2^d)$. In addition, for τ-low d-juntas with $\tau < d$, the running time also decreases from $O(n^{d+O(1)})$ to $O(n^{\tau+O(1)})$.

6. Learning the relevant variables from uniformly distributed noisy samples

The detection of relevant variables works similarly as in the noise-free case. The following modifications to τ-FOURIER$_d$ (Algorithm 1) vaccinate it against noise; the resulting algorithm τ-NOISY-FOURIER$_d$ is presented as Algorithm 2.

First, the noisy version has to obtain some information about the noise parameters. In the variant presented here, it receives the bounds γ_a, γ_b as additional inputs. Next, to ensure that, in line 5 of the algorithm, β is an appropriate measure to decide whether the Fourier coefficient $\hat{f}(I)$ vanishes, we divide the expression given in the noise-free setting by $\gamma_a^{|I|}\cdot\gamma_b$, which is a lower bound for $|1-2p_I|\cdot|1-2\eta|$.

Additionally to the adaptations of the algorithm, the number of examples that have to be drawn increases by a factor of $4\cdot(\gamma_a^{\tau}\cdot\gamma_b)^{-2}$. Furthermore, instead of receiving a noise-free sample, the algorithm now obtains a noisy sample as input. In particular, in line 1 of τ-NOISY-FOURIER$_d$, $x^k = x'^k \oplus a^k$ and $y^k = y'^k \cdot b^k$ for appropriate noise-free data x'^k, y'^k and noise a^k, b^k.

Theorem 6.1. *Let f be a τ-low d-junta and*

$$m \geq 8\cdot\ln(2n/\delta)\cdot 2^{2d}\cdot(\gamma_a^{\tau}\cdot\gamma_b)^{-2}.$$

Then τ-NOISY-FOURIER$_d(S) = \mathrm{rel}(f)$ with probability at least $1-\delta$. Furthermore, τ-NOISY-FOURIER$_d(S)$ runs in time $n^{\tau}\cdot\mathrm{poly}(m,n)$.

Proof. Let $\rho = 2^{-d}$. Algorithm τ-NOISY-FOURIER$_d$ classifies x_i as "relevant" if and only if $|\tilde{f}_S(I)| \geq (1/2)\cdot\gamma_a^{|I|}\cdot\gamma_b\cdot\rho$ for some I of size at most τ with $i\in I$. By Lemma 5.5, for every $I\subseteq[n]$ of size at most τ,

$$|\tilde{f}_S(I) - (1-2p_I)(1-2\eta)\hat{f}(I)| \leq \tfrac{1}{2}\cdot\gamma_a^{\tau}\cdot\gamma_b\cdot\rho \tag{9}$$

with probability at least $1-\delta/n$.

Consider some variable $x_i\in\mathrm{rel}(f)$. Since f is τ-low, there exists an $I\subseteq[n]$ of size at most τ such that $i\in I$ and $\hat{f}(I)\neq 0$. Since $\hat{f}(I)$ is an integer multiple of $2^{-|\mathrm{rel}(f)|}$, $|\hat{f}(I)|\geq 2^{-d}$. In particular, if (9) is satisfied, then

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

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

Algorithm 3 τ-NOISY-LEARN$_d$

```
input S = ((x_1^k, ..., x_n^k), y^k)_{k∈[m]}, P, γ_a, η
γ_b ← |1 − 2η|
R ← τ-NOISY-FOURIER_d(S, γ_a, γ_b)
for I ⊆ R do
    f̃_S(I) ← (1/m) Σ_{k=1}^m χ_I(x^k) · y^k
output hypothesis
```

$$\tau\text{-NOISY-LEARN}_d(x) = \operatorname{sgn} \sum_{I \subseteq R} (1-2p_I)^{-1}(1-2\eta)^{-1} \tilde{f}_S(I)\chi_I(x)$$

Now consider some variable $x_i \notin \operatorname{rel}(f)$. Thus, $\hat{f}(I) = 0$ for all $I \subseteq [n]$ with $i \in I$ by Lemma 4.1. By (9), with probability at least $1 - \delta/n$,

$$|\tilde{f}_S(I)| \leq \frac{1}{2} \cdot \gamma_a^\tau \cdot \gamma_b.$$

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 $n \cdot (\delta/n) = \delta$. □

7. Constructing a hypothesis from uniformly distributed noisy samples

Learning juntas (in the sense of constructing an accurate hypothesis) in the presence of noise proceeds 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 from the algorithm used in the noise-free setting is that we cannot simply read off the truth table from the examples since these may contain inconsistencies (even if not, such a truth table is unlikely to be correct).

Fortunately, we have seen in Section 5.2 how to build a good hypothesis in the presence of attribute noise. The trick is that we do not apply Proposition 5.8 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 any longer, 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 5.9). Moreover, for the subclass $\mathcal{J}_d^n(\tau)$, the time complexity depends on n^τ instead of n^d. Precisely, the algorithm, which we call τ-NOISY-LEARN$_d$, is presented as Algorithm 3.

Theorem 7.1. *Algorithm* τ-NOISY-LEARN$_d$ *exactly learns the class* $\mathcal{J}_d^n(\tau)$ *with confidence* $1 - \delta$

- *from uniformly distributed* (P, η)*-noisy samples of size* $\operatorname{poly}(\log n, 2^d, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$
- *with running time* $n^\tau \cdot \operatorname{poly}(n, 2^d, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$.

Proof. Let $f \in \mathcal{J}_d^n(\tau)$. As we have shown in Theorem 6.1, with probability at least $1 - \delta/2$, τ-NOISY-FOURIER$_d$ successfully infers the relevant variables of f, provided that

$$m \geq 8 \cdot \ln(4n/\delta) \cdot 2^{2d} \cdot (\gamma_a^\tau \cdot \gamma_b)^{-2}.$$

By Proposition 5.8, again with probability at least $1 - \delta/2$, hypothesis h exactly coincides with f. Hence, τ-NOISY-LEARN$_d$ succeeds in exactly learning the target concept with probability at least $1 - \delta$. The claimed sample complexity and running time follow from Theorem 6.1 and Proposition 5.8. □

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

Corollary 7.2. (a) *The class* $\mathcal{J}_d^n$ *can be exactly* (P, η)*-learned with confidence* $1 - \delta$ *from a sample of size* $\operatorname{poly}(\log n, 2^d, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$ *in running time* $n^d \cdot \operatorname{poly}(n, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$.

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

8. Non-uniformly distributed attributes

In this section we show how to generalize our results to product attribute distributions (not to be confused with attribute *noise* distributions). We confine ourselves to presenting results for 1-low concepts only. The more delicate task of studying the general applicability of the methods to τ-low juntas is 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. To avoid pathological cases, we assume that there exists a constant $\gamma_c \in (0, 1/2]$ such that, for all $i \in [n]$, $d_i \in [\gamma_c, 1-\gamma_c]$. The learning algorithm now has access to D-distributed (P,η)-noisy samples (see Definition 3.1). When using methods from the uniform setting, we now approximate 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 [13]. 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 $\chi_I^D(x) = \prod_{i\in I} \chi_i^D(x)$. 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 $\mathcal{F}_D(f)(I) = \langle f, \chi_I^D\rangle_D$. Since we only work with a single distribution D in the following, we also write $\check{f}(I)$ for $\mathcal{F}_D(f)(I)$ (but reserve $\hat{f}(I)$ to stand for the unbiased Fourier coefficient $\mathcal{F}_{U_n}(f)(I)$). It is not difficult to see that Lemma 4.1 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 simply 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 8.1. Its proof relies on explicit calculations of the *biased* Fourier coefficients of the *unbiased* parity functions.

Lemma 8.1. *Let $f : \{0,1\}^n \to \mathbb{R}$ and $I \subseteq [n]$. Then*

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

Before we prove Lemma 8.1, we calculate the values $\langle \chi_I^{U_n}, \chi_J^D\rangle_D$. This may be of independent interest for other applications since these are the entries of the change of basis matrix for converting coordinates with respect to the unbiased basis $(\chi_I^{U_n} \mid I \subseteq [n])$ to coordinates with respect to the D-biased basis $(\chi_I^D \mid I \subseteq [n])$.

Lemma 8.2. *Let $J \subseteq I \subseteq [n]$. Then*

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

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

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

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

Proof of Lemma 8.1. We first show that $\langle \chi_I^{U_n}, \chi_J^D\rangle_D = 0$ for all $J\not\subseteq I$:

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

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

$$\begin{aligned}
\left\langle f, \chi_I^{U_n}\right\rangle_D &= \left\langle f, \sum_{J\subseteq[n]}\left\langle \chi_I^{U_n}, \chi_J^D\right\rangle_D\cdot\chi_J^D\right\rangle_D = \sum_{J\subseteq I}\left\langle f,\chi_J^D\right\rangle_D\cdot\left\langle \chi_I^{U_n},\chi_J^D\right\rangle_D\\
&= \sum_{J\subseteq I}\check{f}(J)\cdot\mathcal{F}_D(\chi_I^{U_n})(J)\\
&= \sum_{J\subsetneq I}\check{f}(J)\cdot\mathcal{F}_D(\chi_I^{U_n})(J) + \check{f}(I)\cdot\mathcal{F}_D(\chi_I^{U_n})(I).
\end{aligned}$$

Hence,

$$\check{f}(I) = \left(\mathcal{F}_D(\chi_I^{U_n})(I)\right)^{-1}\cdot\left(\left\langle f,\chi_I^{U_n}\right\rangle_D - \sum_{J\subsetneq I}\check{f}(J)\cdot\mathcal{F}_D(\chi_I^{U_n})(J)\right).$$

The claim now follows from Lemma 8.2. □

Algorithm 4 NOISY-PRODUCT-FOURIER$_d$

```
input S = ((x_1^k, ..., x_n^k), y^k)_{k∈[m]}, D, P, η, ρ
R ← ∅
φ_0 ← 1/((1−2η)·m) Σ_{k=1}^m y^k
for i = 1 to n do
    φ_i ← (1 − 2d_i) · φ_0
    ψ_i ← 1/((1−2p_i)·(1−2η)·m) Σ_{k=1}^m y^k · χ_i^{U_n}(x^k)
    β_i ← (ψ_i − φ_i)/(2·√(d_i·(1−d_i)))
    if |β| ≥ ρ/2
      then R ← R ∪ {x_i}
output NOISY-PRODUCT-FOURIER_d(S, D, P, η, ρ) = R
```

9. Learning the relevant variables from non-uniformly distributed noisy samples

The threshold to recognize nonzero Fourier coefficients is given by the least absolute value of the considered nonzero coefficients. Thus, we define the *Fourier threshold* $\mathrm{thr}_D(f)$ of f with respect to D by

$$\mathrm{thr}_D(f) = \min\left\{ \left| \check{f}(i) \right| \;\middle|\; x_i \in \mathrm{rel}(f) \right\}. \tag{10}$$

For concepts f that are not 1-low (with respect to D), $\mathrm{thr}_D(f) = 0$. If $D = U_n$ is the uniform distribution, then for 1-low concepts f, $\mathrm{thr}_D(f) \geq 2^{-|\mathrm{rel}(f)|}$.

For the next theorem, we stick to the notation fixed in the beginning of Section 5, except that S is now assumed to be a *D-distributed* (P, η)-noisy sample of size m.

Theorem 9.1. *Let $f : \{0,1\}^n \to \{-1,+1\}$ be a d-junta with $\rho = \mathrm{thr}_D(f) > 0$ and*

$$m \geq 2 \cdot \ln(4n/\delta) \cdot \rho^{-2} \cdot (\gamma_a \cdot \gamma_b)^{-2} \cdot (\gamma_c \cdot (1-\gamma_c))^{-1}.$$

Then

$$\text{NOISY-PRODUCT-FOURIER}_d(S, D, P, \eta, \rho) = \mathrm{rel}(f)$$

with probability at least $1 - \delta$. Furthermore, the algorithm runs in time

$$\mathrm{poly}(n, \log(1/\delta), \gamma_a^{-1}, \gamma_b^{-1}, \gamma_c^{-1}, \rho^{-1}).$$

Proof. The proof is an extension of the proof of Theorem 6.1. By Lemma 8.1,

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

Since $\mathbb{E}_{x\sim D, b\sim\eta}[f(x)\cdot b] = (1-2\eta)\cdot \check{f}(\emptyset)$, it follows analogously to the proof of Lemma 5.5 that with probability at least $\delta/(2n)$,

$$|\phi_i - (1-2d_i)\cdot f(\emptyset)| \leq \sigma_i \cdot \rho/2,$$

provided that

$$m \geq 2 \cdot \ln(4n/\delta) \cdot \frac{4(1-2d_i)^2}{(1-2\eta)^2 \cdot \sigma_i^2 \cdot \rho^2}. \tag{11}$$

Moreover, we have

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

Thus, with probability at least $1 - \delta/(2n)$,

$$\left| \psi_i - \left\langle f, \chi_I^{U_n} \right\rangle_D \right| \leq \sigma_i \cdot \rho/2,$$

provided that

$$m \geq 2 \cdot \ln(4n/\delta) \cdot \frac{4}{(1-2p_i)^2 \cdot (1-2\eta)^2 \cdot \sigma_i^2 \cdot \rho^2}. \tag{12}$$

The number of examples in the claim dominates both numbers given in (11) and (12). Thus, with probability at least $1-\delta/n$,

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

NOISY-PRODUCT-FOURIER$_d$ classifies x_i as "relevant" if and only if $|\beta_i| \geq \rho/2$. If $\check{f}(i) = 0$, then $|\beta_i| < \rho/2$ with probability at least $1-\delta/n$, and if $\check{f}(i) \neq 0$, then $|\beta_i| \geq \rho/2$ with probability at least $1-\delta/n$ (since $\check{f}(i) \geq \rho$ by assumption). Consequently, all variables are classified correctly with probability at least $1-\delta$. □

For monotone concepts, we obtain

Lemma 9.2. *Let $f : \{0,1\}^n \to \{-1,+1\}$ be a monotone Boolean concept. Then*

$$\mathrm{thr}_D(f) \geq 2 \cdot \min_{x_i \in \mathrm{rel}(f)} \sigma_i \cdot \prod_{x_j \in \mathrm{rel}(f)\setminus\{x_i\}} \min\{d_j, 1-d_j\}.$$

In particular,

$$\mathrm{thr}_D(f) \geq 2 \cdot \prod_{x_i \in \mathrm{rel}(f)} \min\{d_i, 1-d_i\}.$$

Proof. Let $x_i \in \mathrm{rel}(f)$. Then

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

where for $J \subseteq [n]$ and $x \in \{0,1\}^J$, we define $D(x) = \prod_{j \in J} d_i^{x_i} \cdot (1-d_i)^{1-x_i}$, 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 $f_{x_i=0} \geq f_{x_i=1}$. If, in addition, 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,

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

We conclude the proof by showing $\sigma_i \geq \min\{d_i, 1-d_i\}$. If $d_i \leq 1/2$, then $\sigma_i = \sqrt{d_i \cdot (1-d_i)} \geq d_i = \min\{d_i, 1-d_i\}$. If $d_i \geq 1/2$, then $\sigma_i \geq 1 - d_i = \min\{d_i, 1-d_i\}$. □

The lemma also holds for unate concepts.

While under the uniform distribution, the parity function χ_I is $|I|$-low but not $(|I|-1)$-low, the situation is entirely different for non-uniform distributions:

Lemma 9.3. *Let $f : \{0,1\}^n \to \{-1,+1\}$ be a parity function, i.e., $f = \chi_I$ for some $I \subseteq [n]$. Then*

$$\mathrm{thr}_D(f) = 2 \cdot \min_{i \in I}\left(\sigma_i \cdot \prod_{j \in I\setminus\{i\}} |1-2d_j|\right).$$

In particular, if D is a non-degenerate θ-bounded product distribution (i.e., for all $i \in [n]$, $|1-2d_i| \geq \theta > 0$, see Definition 3.4), then

$$\mathrm{thr}_D(f) \geq 2 \cdot \gamma_c \cdot \theta^{d-1}. \tag{13}$$

Proof. Let $i \in \mathrm{rel}(f) = I$. By Lemma 8.2,

$$\mathcal{F}_D(\chi_I)(i) = \left\langle \chi_I^{U_n}, \chi_i^D \right\rangle_D = 2\sigma_i \cdot \prod_{j \in I \setminus \{i\}} (1-2d_j),$$

which proves the equation in the claim. To see the inequality (13), note that $\sigma_i \geq \sqrt{\gamma_c(1-\gamma_c)} \geq \gamma_c$ (since $1-\gamma_c > \gamma_c$). □

In particular, if $d_i \notin \{0, \frac{1}{2}, 1\}$ for all $i \in [n]$, then the relevant variables of parity functions can be inferred via the Fourier approach (even in the presence of noise). Furthermore, since the relevant variables already determine the target concept in this case, the learning problem is as easy as the detection of relevant variables.

For the class of monotone d-juntas and the class of parity d-juntas we obtain

Corollary 9.4. (a) *The relevant variables of monotone d-juntas can be exactly learned with confidence $1-\delta$*

- *from D-distributed (P, η)-noisy samples of size*

$$m \geq \mathrm{poly}(\log n, \log(1/\delta), \gamma_a^{-1}, \gamma_b^{-1}, \gamma_c^{-d})$$

- *with running time* $\mathrm{poly}(m, n)$.

(b) *If D is θ-bounded, then the class PAR_d^n of parity d-juntas can be exactly learned with confidence $1-\delta$*

- *from D-distributed (P, η)-noisy samples of size*

$$m \geq \mathrm{poly}(\log n, \log(1/\delta), \gamma_a^{-1}, \gamma_b^{-1}, \gamma_c^{-1}, \theta^{-d})$$

- *with running time* $\mathrm{poly}(m, n)$.

Proof. Part (a) follows from Theorem 9.1 and Lemma 9.2; part (b) follows from Theorem 9.1 and Lemma 9.3. Note that a parity function f is uniquely determined by $\mathrm{rel}(f)$. □

10. Constructing a hypothesis from non-uniformly distributed noisy samples

Next we describe how to construct a hypothesis for general concepts. We use Lemma 8.1 to successively approximate all biased Fourier coefficients level by level. 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

$$\beta_I = \left((1-2p_I)(1-2\eta) \prod_{i \in I} 2\sigma_i \right)^{-1} \cdot \frac{1}{m} \cdot \sum_{k=1}^{m} y^k \chi_I(x^k) - \sum_{J \subsetneq I} \prod_{i \in I \setminus J} \frac{1-2d_i}{2\sigma_i} \beta_J. \tag{14}$$

Finally, we build the hypothesis $h(x) = \mathrm{sgn} \sum_{I \subseteq R} \beta_I \cdot \chi_I^D(x)$.

To ensure that β_I approximates $\check{f}(I)$ well enough, reasonably good approximations of all coefficients $\check{f}(J)$ for $J \subseteq I$ are required. This feedback effect leads to a necessary sample size of $2^{\omega(|\mathrm{rel}(f)|)}$.

Theorem 10.1. *Let $f : \{0,1\}^n \to \{-1,+1\}$ be a d-junta with $\rho = \mathrm{thr}_D(f) > 0$. Then f can be exactly recovered with confidence $1-\delta$ from D-distributed (P, η)-noisy samples of size*

$$m \geq \mathrm{poly}(\log n, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1}, \gamma_c^{-d^2}, \rho^{-1})$$

with running time $\mathrm{poly}(m, n)$.

Note that $\gamma_c^{-1} \geq 2$. For a fixed attribute distribution D, the sample size is polynomial in 2^{d^2}. Before we prove Theorem 10.1, we show that a suitable hypothesis can be built, provided that the set of relevant variables is already known:

Lemma 10.2. *Let S be a D-distributed (P, η)-noisy sample of size*

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

for f. Let β_I as defined in (14). Then with probability at least $1-\delta$, the hypothesis h defined by

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

coincides with f.

Proof. We first prove by induction on $|I|$ that $|\beta_I - \check{f}(I)| \leq \varepsilon$ with probability at least $1-\delta$, provided that

$$m \geq 8 \cdot \ln\left(2^{(|I|^2+|I|)/2}/\delta\right) \cdot \gamma_a^{-2|I|} \cdot \gamma_b^{-2} \cdot \gamma_c^{-(|I|^2+3|I|)} \cdot \varepsilon^{-2}. \tag{15}$$

We have

$$\beta_\emptyset = (1-2p_\emptyset) \cdot (1-2\eta) \cdot \frac{1}{m} \sum_{k=1}^{m} y^k.$$

By the Hoeffding bound (Lemma 2.1), with probability at least $1-\delta$,

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

which is clearly dominated by (15).

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

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

The remainder of the proof is a bit technical, so we provide a brief overview first: we show that, with probability at least $1 - \delta \cdot 2^{-|I|}$, (16) holds, and that, with probability at least $1 - \delta \cdot (1-2^{-|I|})$, (17) holds for all $J \subsetneq I$. Putting these things together, we will obtain that, with probability at least $1-\delta$, $|\beta_I - \check{f}(I)| \leq \varepsilon$.

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

$$\left|\psi_I - \left(\prod_{i \in I} 2\sigma_i\right)^{-1} \cdot \left\langle f, \chi_I^{U_n} \right\rangle_D\right| \leq \varepsilon/2, \tag{16}$$

provided that

$$m \geq 2 \cdot \ln\left(\frac{2 \cdot 2^{|I|}}{\delta}\right) \cdot \frac{4}{(1-2p_I)^2 \cdot (1-2\eta)^2 \cdot \left(\prod_{i \in I} 2\sigma_i\right)^2 \cdot \varepsilon^2}.$$

Again, this is dominated by (15) since $\sigma_i \geq \min\{d_i, 1-d_i\} \geq \gamma_c$ (as we have shown in the end of the proof of Lemma 9.2) and thus $(\prod_{i \in I} 2\sigma_i)^2 \geq (2 \cdot \gamma_c)^{2|I|} \geq \gamma_c^{|I|^2+3|I|}$ (recall that $0 < \gamma_c \leq 1/2$).

Furthermore, by induction hypothesis, we have that for each $J \subsetneq I$, with probability at least $1 - \delta \cdot 2^{-|I|}$,

$$|\beta_J - \check{f}(J)| \leq \gamma_c^{|I|+1} \cdot \varepsilon, \tag{17}$$

provided that

$$m \geq 8 \cdot \ln\left(2^{(|J|^2+|J|)/2} \cdot 2^{|I|}/\delta\right) \cdot \gamma_a^{-2|J|} \cdot \gamma_b^{-2} \cdot \gamma_c^{-(|J|^2+3|J|)} \cdot (\varepsilon \cdot \gamma_c^{|I|+1})^{-2}.$$

Since $|1-2d_i|/\sigma_i \leq 1/\sigma_i \leq \gamma_c^{-1}$, we obtain

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

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

$$\begin{aligned}m &\geq 8\cdot\ln\left(2^{((|I|-1)^2+|I|-1)/2}\cdot 2^{|I|}/\delta\right)\cdot\gamma_a^{-2(|I|-1)}\cdot\gamma_b^{-2}\cdot\gamma_c^{-((|I|-1)^2+3(|I|-1))}\cdot(\gamma_c^{|I|+1}\cdot\varepsilon)^{-2} \\ &= 8\cdot\ln\left(2^{(|I|^2-2|I|+1+|I|-1)/2+|I|}/\delta\right)\cdot\gamma_a^{-2(|I|-1)}\cdot\gamma_b^{-2}\cdot\gamma_c^{-(|I|^2-2|I|+1+3|I|-3+2|I|+2)}\cdot\varepsilon^{-2} \\ &= 8\cdot\ln\left(2^{(|I|^2+|I|)/2}/\delta\right)\cdot\gamma_a^{-2(|I|-1)}\cdot\gamma_b^{-2}\cdot\gamma_c^{-(|I|^2+3|I|)}\cdot\varepsilon^{-2}.\end{aligned}$$

The latter sample bound is again dominated by (15). Finally,

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

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

Now we apply this result to estimate how closely h approximates f. Assume that $|\beta_I - \check{f}(I)| \leq \sqrt{2^{-d}\cdot\varepsilon}$ for all $I \subseteq R$. The standard LMN analysis (see Linial et al. [18]) yields

$$\Pr_{x\sim D}[h(x)\neq f(x)] \leq \sum_{I\subseteq R}(\beta_I - \check{f}(I))^2 \leq 2^d\cdot(2^{-d}\varepsilon) = \varepsilon.$$

Let $\varepsilon = \gamma_c^d/2$. Then, with probability at least $1-\delta$,

$$\Pr_{x\sim D}[h(x)\neq f(x)] \leq \gamma_c^d/2 < \prod_{i\in R}\min\{d_i, 1-d_i\} = \min_{x\in\{0,1\}^R} D(x).$$

This implies $h = f$. Thus, we can request

$$\begin{aligned}m &\geq \text{poly}(\log(2^{d^2}/\delta), \gamma_a^{-d}, \gamma_b^{-1}, \gamma_c^{-d^2}, \gamma_c^{-d}) \\ &= \text{poly}(\log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1}, \gamma_c^{-d^2})\end{aligned}$$

examples to guarantee $h(x) = f(x)$ for all $x \in \{0,1\}^n$ with probability at least $1-\delta$. □

Now we can prove Theorem 10.1:

Proof of Theorem 10.1. By Theorem 9.1, 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 $m \geq \text{poly}(\log n, \log(1/\delta), \gamma_a^{-1}, \gamma_b^{-1}, \gamma_c^{-1}, \rho^{-1})$. By Lemma 10.2, f can be exactly recovered from

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

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

Corollary 10.3. *The class* MON_d^n *of monotone* d*-juntas can be exactly learned with confidence* $1 - \delta$

- *from* D*-distributed* (P, η)*-noisy samples of size*

$$m \geq \mathrm{poly}(\log n, \log(1/\delta), \gamma_a^{-1}, \gamma_b^{-1}, \gamma_c^{-d^2})$$

- *with running time* $\mathrm{poly}(m, n)$.

11. 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 τ-low d-juntas under product attribute and classification noise with rates different from $1/2$. For $\tau = 1$, 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 (7)). In addition, we have shown how to generalize the methods to non-uniformly distributed examples. This has led to efficient learning algorithms for monotone juntas and parity juntas.

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 for uniformly distributed examples since it requires noise-tolerant learning of parity functions. We suspect that non-trivial lower bounds (based on hardness assumptions) can be shown.

Acknowledgment

The authors are grateful to an anonymous referee for helpful suggestions to improve the presentation in Sections 5.2 and 10.

References

[1] Tatsuya Akutsu, Satoru Miyano, Satoru Kuhara, Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function, J. Comput. Biol. 7 (3–4) (2000) 331–343.
[2] Noga Alon, Joel Spencer, The Probabilistic Method, in: Wiley-Intersci. Ser. Discrete Math. Optim., John Wiley and Sons, 1992.
[3] Dana Angluin, Philip D. Laird, Learning from noisy examples, Machine Learning 2 (4) (1988) 343–370.
[4] Raghu Raj Bahadur, A representation of the joint distribution of responses to n dichotomous items, in: Herbert Solomon (Ed.), Studies in Item Analysis and Prediction, Stanford University Press, Stanford, CA, 1961, pp. 158–168.
[5] William Beckner, Inequalities in Fourier analysis, Ann. of Math. (2) 102 (1) (1975) 159–182.
[6] Itai Benjamini, Gil Kalai, Oded Schramm, Noise sensitivity of Boolean functions and applications to percolation, Inst. Hautes Études Sci. Publ. Math. 90 (1999) 5–43.
[7] Anna Bernasconi, Mathematical techniques for the analysis of Boolean functions, Ph.D. Thesis, Università degli Studi di Pisa, Dipartimento di Ricerca in Informatica, March 1998.
[8] Avrim Blum, Pat Langley, Selection of relevant features and examples in machine learning, Artificial Intelligence 97 (1–2) (1997) 245–271.
[9] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, Manfred K. Warmuth, Occam's razor, Inform. Process. Lett. 24 (6) (1987) 377–380.
[10] Aline Bonami, Étude des coefficients de Fourier des fonctions de $l^p(g)$, Ann. Inst. Fourier 20 (2) (1970) 335–402.
[11] Nader H. Bshouty, Jeffrey C. Jackson, Christino Tamon, Uniform-distribution attribute noise learnability, Inform. Comput. 187 (2) (2003) 277–290.
[12] Scott E. Decatur, Rosario Gennaro, On learning from noisy and incomplete examples, in: Proceedings of the Eigth Annual Conference on Computational Learning Theory, COLT 1995, Santa Cruz, California, USA, ACM Press, 1995, pp. 353–360.
[13] Merrick L. Furst, Jeffrey C. Jackson, Sean W. Smith, Improved learning of AC^0 functions, in: Leslie G. Valiant, Manfred K. Warmuth (Eds.), Proceedings of the Fourth Annual Workshop on Computational Learning Theory, COLT 1991, Santa Cruz, California, USA, Morgan Kaufmann, 1991, pp. 317–325.
[14] Sally A. Goldman, Robert H. Sloan, Can PAC learning algorithms tolerate random attribute noise? Algorithmica 14 (1) (1995) 70–84.
[15] Wassily Hoeffding, Probability inequalities for sums of bounded random variables, J. Amer. Statist. Assoc. 58 (1963) 13–30.

[16] Jeff Kahn, Gil Kalai, Nathan Linial, The influence of variables on Boolean functions (extended abstract), in: 29th Annual Symposium on Foundations of Computer Science, FOCS '88, IEEE Computer Society, White Plains, NY, 1988, pp. 68–80.
[17] Mihail N. Kolountzakis, Evangelios Markakis, Aranyak Mehta, Learning symmetric juntas in time $n^{o(k)}$, in: Workshop on Interface between Harmonic Analysis and Number Theory, Marseille, 2005, 2005. Available as Tech. Rep. arXiv:math.CO/0504246 v1 at: http://arxiv.org/abs/math.CO/0504246v1.
[18] Nathan Linial, Yishay Mansour, Noam Nisan, Constant depth circuits, Fourier transform, and learnability, J. ACM 40 (3) (1993) 607–620.
[19] Nick Littlestone, Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm, Machine Learning 2 (4) (1987) 285–318.
[20] Akinobu Miyata, Jun Tarui, Etsuji Tomita, Learning Boolean functions in AC^0 on attribute and classification noise, in: Shai Ben-David, John Case, Akira Maruoka (Eds.), Algorithmic Learning Theory, 15th International Conference, ALT 2004, Padova, Italy, October 2–5, 2004, Proceedings, in: Lecture Notes in Artificial Intelligence, vol. 3244, Springer, 2004, pp. 142–155.
[21] Elchanan Mossel, Ryan W. O'Donnell, On the noise sensitivity of monotone functions, Random Structures Algorithms 23 (3) (2003) 333–350.
[22] Elchanan Mossel, Ryan W. O'Donnell, Rocco A. Servedio, Learning functions of k relevant variables, J. Comput. System Sci. 69 (3) (2004) 421–434.
[23] Ryan W. O'Donnell, Computational applications of noise sensitivity, Ph.D. Thesis, Department of Mathematics, Massachusetts Institute of Technology, June 2003.
[24] Rocco A. Servedio, On learning monotone DNF under product distributions, Inform. Comput. 193 (1) (2004) 57–74.
[25] George Shackelford, Dennis Volper, Learning k-DNF with noise in the attributes, in: Proceedings of the 1988 Workshop on Computational Learning Theory, August 3–5, 1988, MIT, Morgan Kaufmann, 1988, pp. 97–103.
[26] Leslie G. Valiant, A theory of the learnable, Commun. ACM 27 (11) (1984) 1134–1142.