Learning Concepts with Few Unknown Relevant Attributes from Noisy Data

Jan Arpe

Dissertation

Universität zu Lübeck
Institut für Theoretische Informatik

Aus dem Institut für Theoretische Informatik
der Universität zu Lübeck
Direktor: Prof. Dr. math. Rüdiger Reischuk

Learning Concepts with Few Unknown Relevant Attributes from Noisy Data

Inauguraldissertation

zur Erlangung der Doktorwürde
der Universität zu Lübeck
aus der Technisch-Naturwissenschaftlichen Fakultät

vorgelegt von

Jan Arpe

Lübeck, August 2006

Berichterstatter:	Prof. Dr. Rüdiger Reischuk
	Prof. Dr. Hans Ulrich Simon
	Prof. Dr. Georg Schnitger
	Prof. Dr. Thomas Zeugmann
Vorsitz des Prüfungsausschusses:	Prof. Dr. Jürgen Prestin
Tag der mündlichen Prüfung:	8. Dezember 2006
zum Druck genehmigt:	8. Dezember 2006

gezeichnet Prof. Dr. Enno Hartmann
Dekan der Technisch-Naturwissenschaftlichen Fakultät der Universität zu Lübeck

Acknowledgments

In the first place, I would like to thank my *Doktorvater* Rüdiger Reischuk for giving me the opportunity to work and to carry out my research at the Institut für Theoretische Informatik of the Universität zu Lübeck and for his scientific advice and support.

I am grateful to my former office mate Bodo Manthey for innumerable scientific discussions, listening to my ideas, and providing helpful suggestions. Also, I received much valuable advice and moral support from Hagen Völzer, Gerhard Buntrock, Till Tantau, Arfst Nickelsen, Andreas Jakoby, Maciej Liśkiewicz, and Thomas Zeugmann. Special thanks go to Frank Balbach, Markus Hinkelmann, Bodo Manthey, Arfst Nickelsen, Till Tantau, and Hagen Völzer for proofreading parts of this thesis. I want to thank all current and former members of the institute for the pleasant working atmosphere.

At this point, I would also like to thank the two persons who introduced the subject of computational complexity theory to me: Thomas Schwentick and Clemens Lautemann. Sadly, Clemens Lautemann passed away last year. I will remember him as one of the most likable professors I have met.

Completing a doctoral thesis certainly puts anybody into good humor, and one tends to forget about the hard times en route. I would like to express my deepest gratitude to my wife Tatiana and to my parents Birgit and Rolf for their constant support and for being there for me—particularly in the not-so-sunny days. Retrospectively, I also appreciate my brother's nasty questions about what on earth I am actually doing in my job. Thanks for keeping me on the ground, Tim!

And last but not least I would like to thank my daughters Marit and Bente for being there—they have acted as the best possible counterbalance to all that academic business. It is an invaluable gift to live with the best kids and the best wife on earth!

Abstract

In this thesis, we are concerned with two major challenges in computational learning theory: learning in the presence of large amounts of irrelevant information and learning from noisy data. We model the former issue by assuming that the concepts to be learned depend only on few relevant attributes—such concepts are called *juntas*. The latter issue is modeled by a random noise process that affects attribute and classification values. Thereby, we confine ourselves to studying Boolean attributes and classifications. We approach the coincidence of both issues from two different perspectives.

First, we investigate a specific greedy algorithm for finding the relevant attributes of a Boolean concept. This algorithm is very simple and successfully used in practice. We provide a precise characterization of the concepts for which the greedy algorithm is successful. This characterization is based on a property of the Fourier spectrum of the concept under consideration. In addition, we show that the algorithm can tolerate quite general attribute and classification noise.

Second, we design and analyze Fourier-based algorithms with the explicit goal of efficiently learning large classes of juntas from uniformly distributed noise-corrupted examples. It turns out that the Fourier method and an extension of the greedy method are capable of learning exactly the same concept classes with equal efficiency. We extend the Fourier approach to non-uniformly distributed examples and prove that monotone juntas and parity functions with few relevant variables can be efficiently learned in this setting.

Both approaches are inefficient in learning the class of parity juntas from uniformly distributed noisy examples. For this task, we propose an alternative method, which is based on the method of minimizing the disagreements between the learning data and the output hypothesis. While the running time of this method is also very high, it works independently of the input distribution.

Furthermore, we prove lower bounds on the sample size that is necessary for successful learning of parity juntas from noisy data in terms of certain noise parameters.

As a side-product, we prove a characterization of general learnability from noise-affected uniformly distributed examples.

Contents

CHAPTER 1

Introduction and Summary

1.1 Motivation

In this thesis, we deal with topics from computational learning theory. This field studies algorithms that allow computers to learn. Specifically, we are concerned with the design and analysis of efficient learning algorithms that can deal at the same time with large amounts of irrelevant information and with noisy data.

A typical data mining scenario is the following: a sequence of training data points is given, each of which consists of an *attribute vector* and a *classification value*. The latter is generated according to an unknown *target concept*, which is a function that maps attribute vectors to classification values. The goal is to produce a hypothesis that, given only an attribute vector not present in the training data, predicts the correct classification value with high probability.

Often, it is the case that the data points contain huge amounts of irrelevant information, i.e., the classification only depends on a small subset of all attributes. For instance, the attribute vectors may be medical records of patients, and the classification indicates whether the patient has a certain disease. The goal is here to identify a small number of attributes that suffice to be checked in order to make an accurate diagnosis. Thereby, it is important to know the minimum number of medical records that are needed to infer a good prediction rule. As another example, the data points may correspond to aligned biological sequences (such as DNA or protein sequences), and the classification may depend only on a few active base positions. Further examples are the classification of web pages, the evaluation of experimental measurements, or the processing of astronomical data.

When modeling the learning scenario for these examples, it is desirable to take into account that the given data are affected by some kind of noise, e.g., due to imprecise measurements, errors occurring during data submission, fuzzy specifications, or other sources of interference.

From a theoretical perspective, learning in the presence of large amounts of irrelevant information and learning from noisy data are among the most challenging issues of algorithmic learning theory and have attracted considerable interest in the past. In this thesis, we investigate what happens when these worlds collide: if an unknown concept depends only on a small number of attributes, how can we learn the target concept from randomly chosen examples that are corrupted by attribute and classification noise? Here, we confine ourselves to look at Boolean attributes and classifications only.

Specifically, we consider a learning model in which the learning algorithm receives a *sample* that consists of random *noisy examples* of the form

$$(x_1 \oplus \xi_1, \ldots, x_n \oplus \xi_n, f(x_1, \ldots, x_n) \oplus \zeta) \in \{0,1\}^n \times \{0,1\} \ ,$$

where the *attribute vectors* $x = (x_1, \ldots, x_n)$ are distributed according to an *attribute distribution* D, the *attribute noise vectors* $\xi = (\xi_1, \ldots, \xi_n)$ are distributed according to an *attribute noise distribution* P, and the *classification noise bit* ζ is set to 1 with some probability η, called the *classification noise rate*. We assume that the unknown target concept $f : \{0,1\}^n \to \{0,1\}$ depends only on a small but unknown set of variables $x_{i_1}, \ldots, x_{i_d}$, where d is much smaller than n. Such functions are called *d-juntas*.

The goal of researchers in this area has been to design fast attribute-efficient algorithms, meaning that the number of examples needed to learn successfully may depend only logarithmically (or poly-logarithmically) on the number n of all attributes. Concerning the number d of relevant attributes, however, exponential dependence on d is often necessary because of information-theoretic reasons. In addition, the output hypotheses are represented by their truth tables of size 2^d (which in the worst case is the most efficient representation). For restricted classes of juntas, it also makes sense to require polynomial dependence in the *description size* of the target concept (which may be much smaller than 2^d), but as we are mainly interested in the worst case, we exclude this issue from our agenda. Ideally, the running time of the algorithms should be polynomial in n and 2^d (and other natural learning parameters). However, for arbitrary juntas, this goal seems out of reach for the time being and may even be impossible.

Already the case of noise-free data (which becomes a special case of the scenario above by letting $P(0^n) = 1$ and $\eta = 0$) is highly interesting. Mossel, O'Donnell, and Servedio [MOS04] believe that "the problem of efficiently learning k-juntas is the single most important open question in uniform distribution learning." Also, Blum [Blu03] stresses the importance of the problem, offering monetary rewards for partial solutions.

Using an exhaustive search algorithm, one can trivially learn the relevant attributes of an arbitrary d-junta from $O(2^d \cdot \log n)$ noise-free examples in time roughly n^d, independently of the attribute distribution. For the uniform attribute distribution, Mossel et al. [MOS04] have provided an essential improvement of this bound. They achieve a running time of roughly $n^{0.704 \cdot d}$. Their analysis combines a *Fourier-based* and a *parity-based* method. The former method searches for nonzero Fourier coefficients of the target function up to a certain level. The latter method solves a large system of linear equations over the two-element field GF(2) with "parity variables" from an expanded variable space (see also Definition 2.2.2). The Fourier method yields an algorithm for learning the class of monotone d-juntas in time polynomial in n and 2^d.

In the noise-free case, it suffices to correctly identify the relevant attributes since the truth-table of the target concept can be deduced by restricting the examples to the relevant attributes. This approach is impossible in case of noisy data, where we have to seek for more clever methods to construct a suitable hypothesis. Even worse: already the exhaustive search algorithm for finding the relevant attributes fails for noisy data. In addition, as we argue in Chapter 6, there is little hope that the *parity-based* method mentioned above has tractable analogs in the scenario of corrupted data.

Concerning the noise distributions, we impose the following restrictions. First, it is reasonable to require $\eta \neq 1/2$ since otherwise the classifications are turned into purely random bits and learning is made impossible. We assume that there exists a constant $\gamma_b > 0$ such that $|1 - 2\eta| \geq \gamma_b$. Second, we require the attribute noise distribution P to be γ_a*-bounded* for some constant $\gamma_a > 0$. As a proper definition of γ_a-boundedness is a bit technical (see Definition 3.2.6), we only mention here that product distributions with $\Pr[\xi_i = 1] \leq \frac{1}{2} \cdot (1 - \gamma_a)$ for all $i \in \{1, \dots, n\}$ are γ_a-bounded.

We see three directions of research departing from the current state of affairs. The first direction is to design attribute-efficient learning algorithms that improve known learnability results for d-juntas from noisy data in one of the following senses. Either, for a given subclass of the d-juntas, the new algorithms should be *faster* than all previous learning algorithms for that subclass. Or, one may develop attribute-efficient algorithms that learn a concept class that is *strictly larger* than what was known to be learnable before, in the same running time.

The second direction is to derive lower bounds on the running time for attribute-efficient learning of d-juntas, possibly based on complexity theoretic hardness assumptions. Similarly, lower sample bounds for learning algorithms that run in polynomial time are of major interest. However, most such lower bounds seem to be tightly connected to apparently hard open problems in complexity theory.

As the third direction, it is worthwhile to study algorithms that are already present in the literature. Especially, if an algorithm is successfully used in practice, it makes sense to analyze its performance in a theoretical framework. In some situations, this may help to decide a priori whether an algorithm is likely to be successful in a specific application or not. In our view, such analyses are valuable even if the results do not contribute to the first direction of research mentioned above, i.e., if they do not yield new general learnability results. Understanding the capacity of algorithms is a central task in computer science.

1.2 Results of this Thesis

We tackle the challenge of learning in the presence of noise and irrelevant information from two different perspectives, which correspond to the third and to the first research direction mentioned above. First, we study a specific greedy algorithm that is very simple and efficient and already used quite often in practice. We precisely characterize the class of concepts for which this algorithm succeeds with high probability. We first prove the characterization for noise-free data and then show that it also holds if the given data are affected by random attribute and classification noise.

Second, we design an algorithm with the explicit goal of efficiently learning a large class of concepts from noisy data. The only approach that seems to work at all is to use Fourier analysis on the hypercube. This is particularly the case if one does not only want to learn the relevant attributes but construct a hypothesis that is supposed to be correct with high probability.

Finally, we take a look at the bottleneck of learning in the presence of noise and irrelevant information: learning parity functions of a small number of variables from uniformly distributed noisy examples. We describe a new method to learn such parity functions and prove some lower sample bounds with respect to the noise parameter γ_a.

The respective results are detailed in the following subsections.

1.2.1 The Greedy Method

The greedy algorithm infers relevant attributes from random data. The task of constructing an output hypothesis has to be completed by additional methods (such as the Fourier method).

To infer relevant attributes from the random sample S, the key task is to find a minimal set of attributes $R \subseteq \{x_1, \ldots, x_n\}$ such that S admits a consistent hypothesis h that depends only on the variables in R. By the principle of Occam's razor [BEHW87], if the sample size m is $\mathrm{poly}(2^d, \log n)$, with high probability there remains only one such hypothesis—the target concept itself. We reduce the problem of finding such a set R to the Set Cover problem, which is one of the best studied NP-complete problems in complexity theory [GJ79]. The reduction maps the sample S to the following Set Cover instance. The ground set is the set of all pairs $\{k, \ell\}$ with $y^k \neq y^\ell$. A pair $\{k, \ell\}$ may be covered by any attribute x_i such that $x_i^k \neq x_i^\ell$. The goal is to cover the ground set by as few attributes as possible. This reduction opens the door to apply well-known greedy heuristics: the most generic one, which we call Greedy, successively selects the largest remaining set and deletes all covered elements, see Johnson [Joh74] or Chvátal [Chv79].

To characterize the functions for which Greedy outputs the relevant attributes, we introduce the notion of *Fourier-accessibility*. A function is Fourier-accessible if and only if all of its relevant variables can be found by exploring the *Fourier support graph*. This is the subgraph of the n-dimensional hypercube induced by the subsets I with nonzero Fourier coefficients $\hat{f}(I)$. The exploration starts with the empty set. Equivalent conditions are provided in Definition 2.4.6 and Lemma 2.4.7.

The readers who have little or no knowledge about Fourier analysis on the hypercube should think of the Fourier coefficient $\hat{f}(I)$, $I \subseteq \{1, \ldots, n\}$, as a measure for the correlation between the function value $f(x)$ and the parity $\bigoplus_{i \in I} x_i$, where $x \in \{0,1\}^n$ is drawn according to the uniform distribution. Several different views of Fourier coefficients are presented in Section 2.2. Additionally, a view on Fourier analysis on the hypercube from a more structural and mathematically more sophisticated standpoint is briefly described in Section 7.2.

We prove that if the target concept $f\colon \{0,1\}^n \to \{0,1\}$ is a Fourier-accessible d-junta, then Greedy correctly infers all relevant variables of f under the uniform distribution from $\mathrm{poly}(2^d, \log n, \log(1/\delta))$ examples with probability at least $1 - \delta$. On the other hand, we show that if f is *not* Fourier-accessible, then the error probability of Greedy is at least $1 - d^2/(n-d)$, independent of the sample size. In particular, this probability tends to 1 as d is fixed and $n \to \infty$ or as $d \to \infty$ and $n \in \omega(d^2)$. Thus, the average-case analysis of the greedy algorithm results in a dichotomy: for a given concept, *either* the relevant variables are inferred correctly with high probability *or* at least one relevant variable is not detected at all with high probability.

There are simple functions that are not Fourier-accessible (see Example 2.4.8 (d)). This inspires us to extend the concept of Fourier-accessibility to τ-*Fourier-*

accessibility (see Definition 2.4.9). Here, during the exploration of the Fourier support graph, we allow for "jumping" to other connected components of Hamming distance at most $\tau - 1$. We devise an extension of GREEDY that correctly infers the relevant variables of a concept f if and only if f is τ-Fourier-accessible.

In another direction, we generalize our investigation of GREEDY to the scenario of noise-affected data. It turns out that the algorithm is very noise-tolerant. As the only adaptations we have to undertake, we additionally have to provide the algorithm with the number of relevant attributes and to feed it with a higher number of examples. Specifically, we prove that the variant GREEDY$_d$ correctly finds the relevant attributes of Fourier-accessible concepts f, provided that the number of given examples is polynomial in 2^d, $\log n$, $\log(1/\delta)$, γ_a^{-d}, and γ_b^{-1} (see Theorem 4.5.3).

In addition to GREEDY, we investigate an even simpler greedy strategy: instead of deleting the covered elements in each round and iteratively computing the resulting sizes of the remaining sets, GREEDY RANKING simply ranks the sets once in the beginning and then selects the largest ones until all ground elements are covered. No set sizes are recalculated. We prove that this variant learns the relevant attributes of a function f if and only if f is 1-*low*. In general, a function is τ-*low* if for each relevant variable x_i there exists an $I \subseteq \{1, \ldots, n\}$ such that $i \in I$, $|I| \leq \tau$, and the Fourier coefficient $\hat{f}(I)$ is nonzero. In particular, every τ-low function is also τ-Fourier accessible, and the notions coincide for symmetric concepts. Analogously to the results obtained for GREEDY, we generalize GREEDY RANKING to cope exactly with the class of τ-low concepts. Furthermore, we prove that GREEDY RANKING is as robust against noise as GREEDY is.

The main techniques that we employ to derive our results for the greedy algorithm are as follows. First, we express the size of certain sets (related to the SET COVER instance) by means of the sizes of certain other sets that correspond to additionally introduced parity variables $x_I = \bigoplus_{i \in I} x_i$ (Lemma 4.2.1). This way of computing the set size can be seen as a special counting technique that is similar to the inclusion-exclusion principle. Second, we derive a Chernoff style deviation bound for certain set sizes (Lemma 4.2.2). Finally, we find a connection between these expected set sizes and corresponding Fourier coefficients (Lemma 4.2.3).

On the one hand, greedy algorithms have been studied by many researchers. On the other hand, there is long tradition of using properties of the Fourier spectrum of Boolean functions in the design of learning algorithms. The conceptual novelty of our analysis lies in the fact that the GREEDY algorithm does *not* exploit any properties of the Fourier spectrum explicitly. Nonetheless, we show that Fourier-accessibility is necessary and sufficient for this algorithm to work successfully. Thus, we obtain a purely analytical characterization for the correctness set of a nontrivial greedy algorithm.

1.2.2 The Fourier Method

While the characterization of the concepts for which the simple greedy algorithm is successful is not at all obvious (at least not to us), design and analysis of the Fourier-based algorithms go hand in hand. As a consequence, the results for the latter algorithms are easier to understand from a conceptual point of view. The focus in our exposition is therefore on explaining which problems occur when trying to transfer methods from the noise-free scenario to the noisy case.

Moreover, while we do not know how to analyze Greedy in case of non-uniformly distributed attributes, the use of Fourier analysis allows us to extend the Fourier method (with some additional effort) to this setting, at least for product distributions. Finally, there is another advantage of the Fourier method: it naturally leads to the construction of a hypothesis (as opposed to only detecting the relevant attributes).

Concerning the task of designing an algorithm for learning juntas from noisy data, we show that the class of τ-low d-juntas is exactly learnable from $m = \mathrm{poly}(\log n, 2^d, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$ uniformly-distributed noisy examples in time $n^\tau \cdot \mathrm{poly}(m, n, 2^d)$. As a main application, the class of monotone d-juntas, for which $\tau = 1$, is learnable from uniformly distributed noisy examples in time $\mathrm{poly}(n, 2^d, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$. Similarly to the Fourier method described by Mossel et al. [MOS04], the idea is to find the relevant attributes by approximating all Fourier coefficients $\hat{f}(I)$ for $|I| \le \tau$, albeit in our setting from highly disturbed data.

As we have already mentioned, finding the relevant variables does not suffice to build a suitable hypothesis. Instead, we restrict the examples to the relevant variables and apply a learning algorithm for arbitrary concepts. The 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 [LMN93] and extended to the noisy scenario by Bshouty, Jackson, and Tamon [BJT03]. 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 poly-logarithmically on n (and exponentially on d).

We extend our methods to certain non-uniform attribute distributions. More precisely, we assume that the attribute distribution D is a product distribution, i.e., in each example, each attribute x_i is independently set to 1 with some probability d_i. To exclude pathological cases, we impose the restriction that

$d_1, \ldots, d_n \in [\gamma_c, 1 - \gamma_c]$ for some $\gamma_c > 0$. Furthermore, we assume that there exists a $\rho > 0$ such that for all relevant variables x_i, $|\hat{f}(i)| \geq \rho$. We show that in this setting, the relevant attributes of d-juntas are learnable from $m = \text{poly}(\log n, 2^d, \log(1/\delta), \rho^{-1}, \gamma_a^{-1}, \gamma_b^{-1}, \gamma_c^{-1})$ noisy examples in time $\text{poly}(m, n)$. In particular, we show that for monotone d-juntas, we can choose $\rho = 2\gamma_c^d$, and for parity functions of up to d variables, we can choose $\rho = 2\gamma_c\theta^{d-1}$, provided that $|1 - 2d_i| \geq \theta > 0$ for all $i \in \{1, \ldots, n\}$. For the uniform attribute distribution, i.e., for the case $d_1 = \ldots = d_n = 1/2$, the latter class is likely to be *not* efficiently learnable from noisy data, as we argue in Chapter 6.

To output a hypothesis that is correct with probability at least $1 - \delta$, we need $m = \text{poly}(\log n, 2^{d^2}, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1}, \rho^{-1})$ noisy examples and a running time of $\text{poly}(m, n)$, provided that $\gamma_c \geq 0.2764$. It turns out that the extension is not as straightforward as one might first think. 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; in order to approximate a biased Fourier coefficient $\hat{f}(I)$, $I \subseteq [n]$, one already has to have good approximations to all coefficients $\hat{f}(J)$, $J \subsetneq I$.

Although the greedy algorithms and the Fourier method are based on totally different ideas, it turns out that they are successful for essentially the same concept classes. For instance, GREEDY RANKING and 1-FOURIER learn the relevant attributes of the same class of concepts, namely the 1-low concepts, in polynomial time. Similarly, τ-GREEDY RANKING and τ-FOURIER both learn the relevant attributes of τ-low concepts in time roughly n^τ. Both methods are attribute-efficient. Moreover, slight modifications of the Fourier algorithms are also capable of learning Fourier-accessible functions (see also Chapter 7).

Historically, we have constructed the Fourier-based algorithms *before* we had found the Fourier-theoretic analysis of the greedy algorithm and its extensions. Prior to the development of the latter insight, the hope was that the combinatorial greedy approach might be able to deal with concepts for which the Fourier approach fails (or takes $n^{\omega(1)}$ steps). Quite the contrary, it turned out that both methods are equally powerful.

1.2.3 Further Results

We introduce the *noise operator* T_P which maps a function $f : \{0,1\}^n \to \mathbb{R}$ to $T_P(f) : \{0,1\}^n \to \mathbb{R}$. For $x \in \{0,1\}^n$, the value $T_P(f)(x)$ is the expectation of $f(x \oplus \xi)$, where ξ is drawn according to the attribute noise distribution P.

This operator is helpful for simplifying the proofs of some results of Bshouty et al. [BJT03] by embedding them into a more structural framework. Specifically, their *noisy distance* $\Delta(f, g)$ between two Boolean functions f and g is equal to $\|T_P(f-g)\|_1$, whereas a related measure they introduce turns out to be equal to $\|T_P(f-g)\|_2^2$. We extend a positive learning result of Bshouty et al. by providing a more general condition under which their "LMN-style" learning algorithm works (Theorem 3.6.3). Furthermore, we complement the results of Bshouty et al. by proving a general learnability characterization: a concept class $\mathcal{C}$ is learnable with accuracy parameter ϵ from noisy examples if and only if the noisy distance between any pair of ϵ-far concepts in $\mathcal{C}$ is positive. This characterization yields positive and negative results for general learnability from noisy data. As an example, we prove that without restricting the attribute noise distributions to be γ_a-bounded, noise-tolerant learning is impossible in general: as we show in Theorem 3.7.3, there is a simple concept class that is impossible to learn under the (a priori known) noise distribution P constructed in Example 3.2.7. This shows that our results cannot be extended to arbitrary noise distributions. On the other hand, we show that any concept class is in principle learnable under any γ_a-bounded attribute noise distribution.

The Fourier method and the extended greedy method both need roughly n^d steps to learn the relevant variables of parity functions with at most d relevant variables. Thus, different methods are needed for parity functions.

After reviewing the current state of affairs for parity junta learning in Section 6.1, we propose an alternative method for learning parities of few variables under attribute and classification noise: it first reduces the setting to pure classification noise and then uses the method of minimizing the number of disagreements between the hypothesis and the noisy input data, which has been proposed by Angluin and Laird [AL88]. Unfortunately, the complexity of disagreement minimization seems to be no lower than the complexity of the Fourier-based method. Some complexity theoretic issues are discussed in Section 6.1.

We show that the success of the latter approach is closely related to the *noise stability* introduced in Section 3.4. For the class of all parities of up to d attributes, the noisy distance is always larger than the noise stability: while the noise stability is equal to the absolute value of the smallest eigenvalue λ_I, $|I| \leq d$, of the noise operator, the noisy distance is equal to half the absolute value of the second smallest such eigenvalue. We present an example in which the disagreement minimization fails although the concept class under consideration is learnable in principle.

For γ_a-bounded attribute noise distributions with $\gamma_a > 0$, the noise stability (and hence also the noisy distance) of the class of parity juntas are always positive. Exploiting a lower bound due to Bshouty et al. [BJT03], we show that any learning algorithm for the class of d-juntas on n attributes needs $\Omega(\gamma_a^{-d})$ examples. In this respect, the results for the greedy and the Fourier algorithms are optimal.

1.3 Related Work

1.3.1 Learning from Noise-free Data

The general learning model considered in this thesis is a variant of Valiant's PAC learning model [Val84]. In most of our applications, this variant fixes the attribute distribution (as opposed to the original distribution-free framework).

There is a long tradition of relating algorithmic learning problems to spectral properties of Boolean functions, see, e.g., Linial, Mansour, and Nisan [LMN93], Kushilevitz and Mansour [KM93], Mansour [Man94], Blum et al. [BFJ$^+$94], and Bshouty and Tamon [BT96]. Specifically, as we have mentioned already, Mossel, O'Donnell, and Servedio [MOS04] have combined spectral and algebraic methods to learn the class of all n-ary d-juntas in roughly $n^{0.704 \cdot d}$ steps. For symmetric juntas (i.e., juntas invariant under permutations of the relevant attributes), this has been improved to $n^{(3/31)d}$ by Lipton et al. [LMMV05] and subsequently to $n^{O(d/\log(d))}$ by Kolountzakis, Markakis, and Mehta [KMM05]. Recently, Köbler and Lindner [KL06] have provided a survey article on learning via the Fourier transform.

Fourier analysis on the hypercube has been studied by many researchers, both under uniform distribution (see, e.g., Lechner [Lec71], Bernasconi [Ber98], and Štefankovič [Šte00]) and under non-uniform distribution (see Bahadur [Bah61], Furst, Jackson, and Smith [FJS91], Bshouty and Tamon [BT96], and Servedio [Ser04]). More information on the Fourier transform of locally compact Abelian groups can be found in Loomis [Loo53, Chapter VII] and Katznelson [Kat04, Chapter VII]. Terras [Ter99] and Štefankovič [Šte00] provide a lot of applications of Fourier analysis on finite groups.

In data mining, the mere goal of determining the relevant attributes is also known as (a simple form of) *(relevant) feature (subset) selection* or *dimension reduction.* It serves as a preprocessing before applying the actual induction scheme, which may then concentrate on producing a suitable hypothesis from a much smaller amount of data. Blum and Langley [BL97] distinguish between *embedded approaches* (such as Littlestone's Winnow [Lit87]), *filter approaches* (such as the greedy algorithm considered in this thesis), and *wrapper approaches* to relevant feature selection, see also John, Kohavi, and Pfleger [JKP94]. Preferring hypotheses that depend on fewer attributes has been called the *min-features bias* by Almuallim and Dietterich [AD94], who also remark that such functions are *semantically* simpler in that they consider fewer aspects of the data.

Inferring relevant attributes is also related to the well-studied problem of finding *association rules* (also called *functional dependencies* or *functional relations*) (e.g., see Agrawal, Imelienski, and Swami [AIS93] and Mannila and Räihä [MR92]). In the variant considered in this paper, the *target attribute* Y is fixed (and considered as a *classification*) as by Akutsu et al. [AB96, AMK03].

Efficient inference of relevant attributes is applied in computational biology [AMK00], investigation of chemical structures [AB96], data mining in corporate and scientific records [BL97]. Additionally, Blum and Langley [BL97] have pointed out that the "internet has put a huge volume of low-quality information at the easy access of learning systems." A variety of further applications can be found in the literature, see, e.g., Littlestone [Lit87], Almuallim and Dietterich [AD94], and Blum and Langley [BL97].

For relevant feature selection, the reduction to SET COVER and the greedy approach have been proposed independently by Almuallim and Dietterich [AD94] and Akutsu and Bao [AB96]. Experimental results have been obtained for artificially generated instances as well as for real-world data from various areas, see Almuallim and Dietterich [AD94], Akutsu, Miyano, and Kuhara [AMK00, AMK03], and Boros et al. [BHI$^+$03]. Akutsu et al. [AMK03] have shown how to implement GREEDY such that its running time is only $O(mnd)$, where m denotes the number of given examples, n denotes the total number of attributes, and d denotes the number of relevant attributes.

For uniformly distributed attributes, Akutsu et al. [AMK03] have proved that with high probability, GREEDY successfully infers the relevant variables for the concept class of conjunctions of attributes or their negations (i.e., Boolean monomials) and that a small sample size polynomial in 2^d and $\log n$ already suffices. Fukagawa and Akutsu [FA05] have extended this result to functions f that are unbalanced with respect to all of their relevant variables (i.e., for x uniformly chosen at random, $\Pr[f(x) = 1|x_i = 0] \neq \Pr[f(x) = 1|x_i = 1]$ for each relevant x_i). This condition is equivalent to our property of 1-lowness. Since there are concepts that are Fourier-accessible but not 1-low (see Example 2.4.8 (b)), we improve Fukagawa and Akutsu's positive learning result for the greedy algorithm to a strictly larger class of concepts. Furthermore, we are not aware of any negative results for the greedy algorithm prior to ours.

A function that has $\hat{f}(i) = 0$ for all $i \in [n]$ has been called *difficult* by Rosell et al. [RHRP05] since greedy tree learners cannot find the relevant variables by computing the information gain in this case. They propose to use the method of *skewing* to artificially derive an instance that is distributed differently from the original input instance.

Restricted classes of concepts with few relevant attributes have been studied by Haussler [Hau88], Littlestone [Lit87], Blum, Chalasani, and Jackson [BCJ93], Valiant [Val99], Servedio [Ser05], and others.

Some authors consider the *mistake-bounded* learning model. In this on-line setting, one tries to minimize the number of examples for which the current hypothesis turns out to be wrong. There are several ways known how to convert on-line algorithms with low mistake bounds into efficient PAC learning algorithms (see Angluin [Ang87], Angluin and Laird [AL88], and Littlestone [Lit89]).

In some other learning models, the complexity of learning juntas is settled. As Mossel et al. [MOS04] mention, learning the class of all n-ary d-juntas from membership queries is easy, whereas the same task using statistical queries only is provably hard [Kea98]. An algorithm that uses membership queries can ask for the classification bits of specific points, whereas an algorithm that uses statistical queries can ask for the expected value of arbitrary $\{0,1\}$-valued functions $s(x, f(x))$ up to a certain accuracy. Statistical query algorithms can be easily converted into PAC learning algorithms. In addition, statistical query algorithms work in the presence of classification noise.

Also, for distribution-free PAC learning, several hardness results are known, both for general settings [PV88] and for restricted classes of juntas [BCJ93, Ser05].

1.3.2 Learning from Noisy Data

Angluin and Laird [AL88] 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 [SV88] and later by Decatur and Gennaro [DG95]. Bshouty, Jackson, and Tamon [BJT03] introduced the notion of *noisy distance* between concepts and showed how this quantity relates to learning from uniformly distributed examples in the presence of attribute and classification noise. If the noise distribution can be arbitrary and is *unknown* to the learning algorithm, then learning nontrivial classes is impossible, see Bshouty et al. [BJT03]. Goldman and Sloan [GS95] have shown that even if the attribute noise distribution is restricted to be a product distribution, then all noise rates must be bounded by twice the accuracy parameter (which is 2^{-d} for exactly learning d-juntas). However, Bshouty et al. [BJT03] have shown that it suffices to approximate certain noise parameters to obtain positive results. Miyata, Tarui, and Tomita [MTT04] proved that if one restricts all noise rates to be equal to the same value, then AC^0 can be learned in quasi-polynomial time without any knowledge of this value. For our purposes, when dealing with γ_a-bounded attribute noise distributions and a classification noise rate η that satisfies $|1 - 2\eta| \geq \gamma_b > 0$, it suffices to have knowledge of the bounds γ_a, γ_b to infer the relevant variables. To construct an output hypothesis, however, we require the noise distribution to be exactly known to the algorithm.

The *noise operator* T_P is a generalization of the *Bonami-Beckner operator* T_ρ [Bon70, Bec75], formally introduced, e.g., by O'Donnell [O'D03] and previously studied in several contexts, see, e.g., Kahn, Kalai, and Linial [KKL88], Benjamini, Kalai, and Schramm [BKS99], and Mossel and O'Donnell [MO03]. Our general version appears implicitly in the work of Bshouty et al. [BJT03]. As mentioned above, their main measures turn out to be related to norms of T_P applied to the difference of concepts.

The more severe model of *malicious errors*, in which an adversary can arbitrarily influence the classification bits of a randomly chosen fraction of the examples, has been studied by Kearns and Li [KL93]. Their main result is that distribution-free learning with accuracy $1-\epsilon$ in this model is only possible if the error rate is bounded by $\epsilon/(1+\epsilon)$.

1.3.3 Greedy Algorithms

For many application areas, greedy strategies are natural and efficient heuristics. In some cases, such as for simple scheduling problems, it has been shown that greedy strategies actually find a global optimum. To prove such a property, several different proof techniques have been developed (see, e.g., Kleinberg and Tardos [KT05, Chapter 4]). These techniques include *exchange arguments* and the use of *matroid theory*. Specifically, Rado [Rad42] and Edmonds [Edm71] have shown that a generic optimization problem is solved exactly by the greedy algorithm if and only if the structure underlying the search space is a matroid.

For the vast majority of optimization problems, however, greedy heuristics do not always achieve optimal solutions. In such cases, the behavior of greedy algorithms is hardly understood. The question, "What is the subset of the input space for which a greedy algorithm guarantees optimality?" has rarely been answered. One notable exception is the characterization of transportation problems using the *Monge property* by Shamir and Dietrich [SD90] (the property is named after Gaspard Monge who found a similar property in 1781 [Mon81]). Among the many extensions of the concept of matroids (to characterize the class of functions to be optimized for which the greedy algorithm works) we mention the work by Vince [Vin02].

Sometimes one can at least show that a specific greedy algorithm achieves a certain nontrivial approximation ratio. This, for example, holds for the Set Cover problem with a logarithmic approximation factor (see Johnson [Joh74], Chvátal [Chv79], or Slavík [Sla96]). Under the assumption that NP is not contained in a quasi-polynomial time complexity class, this result has been proven to be best possible by Feige [Fei98]. Other heuristics for Set Cover are based on linear programming, see Hochbaum [Hoc82] and references therein.

1.4 How this Thesis Evolved

This thesis has been prepared in the course of the DFG research projects Re 672/3 “Average and Precision Complexity” and Re 672/4 “Robust Inference and Compression”. A major goal of these projects has been to obtain positive robustness results for discrete algorithms. For this purpose, discrete scenarios have been sought in which non-trivial methods achieve good results despite the presence of erroneous data. It turned out that one such suitable setting is provided by algorithmic learning of Boolean functions in the presence of attribute and classification noise, with the additional constraint that only few relevant information is given. On the one hand, we have proved a high fault-tolerance for a simple greedy algorithm. On the other hand, we have designed an algorithm that can cope with a large class of target concepts, based on abstract Fourier analysis on the hypercube (again under the constraints of noisy data and much irrelevant information).

Parts of this work have been presented at the *14th International Conference on Algorithmic Learning Theory* [AR03] as well as at the *Third International Conference on Theory and Applications of Models of Computation* [AR06].

1.5 Structure of this Thesis

In Chapter 2, we introduce notation and definitions used in subsequent chapters. The introduction of the learning and the noise models, however, are deferred to Chapter 3. That chapter also contains the introduction of the noise operator and related concepts as well as basic results concerning the approximation of Fourier coefficients and upper and lower bounds for learning from noisy data. In addition, it contains the characterization of general learnability from noisy examples. Chapter 4 contains the results about the greedy algorithm and its variants. The Fourier method is studied in Chapter 5. In Chapter 6, the special problem of learning parity juntas from noisy data is discussed and some lower bounds are proved. Possible extensions of the algorithms and their analyses, a remark on the Fourier transform from the viewpoint of group representation theory, and open problems are presented in Chapter 7.

To facilitate the navigation in this thesis, a List of Algorithms, a List of Definitions, and a List of Symbols are included at the end of the main text.

CHAPTER 2

Preliminaries

This chapter is organized as follows. In Section 2.1, we provide basic notation and definitions. We establish Fourier analysis on the hypercube, the central tool for most of our results, in Section 2.2. Furthermore, in Section 2.3, we introduce the main objects under consideration in this thesis: *juntas*. In Section 2.4, the concepts of *lowness* and *Fourier-accessibility* are defined.

2.1 General Notation

For $n \in \mathbb{N}$, let $[n] = \{1, \dots, n\}$. Viewing elements of $\{0,1\}^n$ as binary strings of length n, we sometimes write $x_1 \dots x_n$ for $(x_1, \dots, x_n) \in \{0,1\}^n$. Furthermore, we denote by 0^n (1^n) the binary vectors with 0 (1) in all positions. By a *Boolean function*, we mean a function $f : \{0,1\}^n \to \Omega$, where $\Omega = \{0,1\}$ or $\Omega = \{-1,+1\}$, depending on how we choose to represent Boolean function values. If we choose $\Omega = \{0,1\}$, then 0 is interpreted as *false* and 1 is interpreted as *true*, whereas if we choose $\Omega = \{-1,+1\}$, then $+1$ is interpreted as *false* and -1 is interpreted as *true*. In particular, the transformations preserving these interpretations are given by

- $x \mapsto 1 - 2x$ (or $x \mapsto (-1)^x$) from $\{0,1\}$ to $\{-1,+1\}$ and
- $x \mapsto \frac{1}{2}(1 - x)$ from $\{-1,+1\}$ to $\{0,1\}$.

A Boolean function is also called a *concept*. A *concept class* $\mathcal{C}$ is a set of concepts $f : \{0,1\}^n \to \Omega$.

We often identify the set $\{0,1\}^n$ with the power set $\mathcal{P}([n])$ of $[n]$: a vector $x \in \{0,1\}^n$ corresponds to the set $\{i \in [n] \mid x_i = 1\}$. In this context, we also sometimes identify a variable x_i with its index i. Throughout this thesis, we use the terms *variable* and *attribute* interchangeably. We write

$$e_i = (0, \dots, 0, 1, 0, \dots, 0) \in \{0,1\}^n \ ,$$

where the 1 is located at position i. This corresponds to the singleton set $\{i\}$.

In general, for arbitrary sets A and B, we denote by B^A the set of all mappings from A to B.

The binary *exclusive or* operation $\oplus : \{0,1\} \times \{0,1\} \to \{0,1\}$ is defined by $x \oplus y = x + y \mod 2$ and extended componentwise to $\oplus : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^n$. In terms of subsets of $[n]$, this corresponds to taking the *symmetric difference*: for $A, B \subseteq [n]$, we define $A \triangle B = (A \cup B) \setminus (A \cap B)$. In the $\{-1,+1\}$-domain, the *exclusive or* of two bits corresponds to *multiplication*.

A function $f : \{0,1\}^n \to \{0,1\}$ is called a *parity function* or simply a *parity* if there exists $I \subseteq [n]$ such that $f(x) = \bigoplus_{i \in I} x_i$ for all $x \in \{0,1\}^n$. The class of all n-ary parities is denoted by PAR^n. The corresponding functions that map to $\{-1,+1\}$ are also referred to as parity functions.

A concept $f : \{0,1\}^n \to \Omega$ is *monotone* if

- for all $x, y \in \{0,1\}^n$, $x \le y$ implies $f(x) \le f(y)$, or
- for all $x, y \in \{0,1\}^n$, $x \le y$ implies $f(x) \ge f(y)$,

where $x \le y$ means that $x_i \le y_i$ for all $i \in [n]$. The second condition is sometimes called *anti-monotone* in the literature. The class of all n-ary monotone concepts is denoted by MON^n. For $i \in [n]$, f is called *monotone in x_i* if

- for all $x \in \{0,1\}^n$ with $x_i = 0$, $f(x) \le f(x \oplus e_i)$, or
- for all $x \in \{0,1\}^n$ with $x_i = 0$, $f(x) \ge f(x \oplus e_i)$.

Finally, f is called *locally monotone* if it is monotone in each variable x_i, $i \in [n]$.

A concept $f : \{0,1\}^n \to \Omega$ is *symmetric* if permuting the variables does not affect the function value, i.e., if for all $x \in \{0,1\}^n$ and all permutations $\pi : [n] \to [n]$,

$$f(x_{\pi(1)}, \dots, x_{\pi(n)}) = f(x_1, \dots, x_n) \ .$$

Equivalently, f is symmetric if and only if $f(x)$ only depends on the *Hamming weight* $|x|$ of x, which is defined by $|x| = |\{i \in [n] \mid x_i = 1\}|$. Interpreting x as a subset of $[n]$, $|x|$ is equal to the set size of x.

Next we introduce *restrictions* of Boolean vectors and functions. In this context, it is useful to think of a Boolean vector $x \in \{0,1\}^n$ as an *assignment*

$x : [n] \to \{0,1\}$ (thus identifying $\{0,1\}^n$ and $\{0,1\}^{[n]}$). A *partial assignment* a is then a mapping $a : I \to \{0,1\}$ (equivalently, $a \in \{0,1\}^I$) for some $I \subseteq [n]$. For two partial assignments $a : I \to \{0,1\}$ and $b : J \to \{0,1\}$ with $I \cap J = \emptyset$, we define the *join*

$$a \sqcup b : I \cup J \to \{0,1\}$$

in the natural way. We denote by $x|_I : I \to \{0,1\}$ the *restriction* of the mapping $x : [n] \to \{0,1\}$ to some subset $I \subseteq [n]$. In this view, a Boolean function $f : \{0,1\}^n \to \Omega$, interpreted as $f : \{0,1\}^{[n]} \to \Omega$, maps assignments to Ω. Given a partial assignment $a : I \to \{0,1\}$, let the *restriction* of f to a be defined by $f_a : \{0,1\}^{[n]\setminus I} \to \Omega$, $f_a(x) = f(a \sqcup x)$. If $I = \{i\}$, then we also write $f_{x_i=a(i)}$ for f_a.

For $n \in \mathbb{N}$ and $d \in \{0,\ldots,n\}$, the *Hamming sphere of radius d* in $\{0,1\}^n$ is the set of points of Hamming weight at most d. Let $V(n,d)$ denote the size of this sphere, i.e.,

$$V(n,d) = \sum_{i=0}^{d} \binom{n}{i} . \tag{2.1}$$

Equivalently, $V(n,d)$ is equal to the number of subsets of $[n]$ of size at most d. Although many sophisticated and essentially tight upper and lower bounds on $V(n,d)$ are available in the literature, we only need very coarse (but therefore, simple) bounds provided by the following lemma, which is straightforward to prove.

Lemma 2.1.1. *For all $n, d \in \mathbb{N}$ with $n \geq 2$ and $2 \leq d \leq n$,*

$$\left(\frac{n-d}{d}\right)^d \leq V(n,d) \leq n^d .$$

We say that a number N that depends on certain parameters $t_1,\ldots,t_k \in \mathbb{R}$ is of size $\mathrm{poly}(t_1,\ldots,t_k)$ if there is a k-ary real polynomial p such that $N \leq p(t_1,\ldots,t_k)$. For real intervals, we use the standard notation, so e.g., $[0,1] = \{x \in \mathbb{R} \mid 0 \leq x \leq 1\}$ and $[0,\frac{1}{2}) = \{x \in \mathbb{R} \mid 0 \leq x < \frac{1}{2}\}$. The sign function $\mathrm{sgn} : \mathbb{R} \to \{-1,+1\}$ is defined by $\mathrm{sgn}(x) = -1$ if $x < 0$ and $\mathrm{sgn}(x) = +1$ if $x \geq 0$. In particular, we define $\mathrm{sgn}(0) = +1$ for technical reasons. The functions log and ln denote the binary and the natural logarithm, respectively. The function exp denotes the exponential function with base e, the Euler constant.

We denote probabilities by Pr, expected values by $\mathbb{E}$, and variances by Var. Furthermore, for a probability distribution $D : \{0,1\}^n \to [0,1]$, we write $\Pr_{x\sim D}$ if the probability is taken over all x distributed according to D. If $n = 1$ and $\Pr[x = 1] = \eta \in [0,1]$, then we write $x \sim \eta$. We use the same notation for the case that $x \in \{-1,+1\}$: here we indicate by $x \sim \eta$ that $\Pr[x = -1] = \eta$.

Similarly, we use the notation $\mathbb{E}_{x\sim D}$ and $\mathrm{Var}_{x\sim D}$. The uniform distribution on $\{0,1\}^n$ is denoted by U_n, i.e., $U_n(x) = 2^{-n}$ for all $x \in \{0,1\}^n$. A probability distribution is *degenerate* if it assigns zero probability to some element, otherwise it is *non-degenerate*.

If $x \in \{0,1\}^n$ is drawn according to a distribution $D : \{0,1\}^n \to [0,1]$, then a function $f : \{0,1\}^n \to \mathbb{R}$ may also be considered as a real-valued random variable. Thus, if D is clear from the context, we use the notation $\Pr[f = b] = \Pr_{x\sim D}[f(x) = b]$ for $b \in \mathbb{R}$, $\mathbb{E}[f] = \mathbb{E}_{x\sim D}[f(x)]$, and

$$\mathrm{Var}[f] = \mathrm{Var}_{x\sim D}[f(x)] = \mathbb{E}_{x\sim D}[(f(x) - \mathbb{E}[f])^2] \ .$$

If $f : \{0,1\}^n \to \{0,1\}$, then $\mathbb{E}[f] = \Pr[f = 1]$ and

$$\mathrm{Var}[f] = \Pr[f = 0] \cdot \Pr[f = 1] \ .$$

If $f : \{0,1\}^n \to \{-1,+1\}$, then

$$\mathbb{E}[f] = \Pr[f = 1] - \Pr[f = -1] = 1 - 2\Pr[f = -1]$$

and

$$\mathrm{Var}[f] = \Pr[f = +1] \cdot \Pr[f = -1] \ .$$

2.2 Fourier Analysis on the Hypercube

Let the hypercube $\{0,1\}^n$ be equipped with a non-degenerate probability distribution $D : \{0,1\}^n \to [0,1]$. The natural inner product

$$\langle f, g\rangle_D = \mathbb{E}_{x\sim D}[f(x) \cdot g(x)] \tag{2.2}$$

for $f, g : \{0,1\}^n \to \mathbb{R}$ turns the set $\mathbb{R}^{\{0,1\}^n}$ of real-valued functions on the hypercube into a real Hilbert space of dimension 2^n. The norm induced by the inner product is

$$\|f\|_D = \sqrt{\langle f, f\rangle_D} \ . \tag{2.3}$$

Given an orthonormal basis $(b_t \mid t \in T)$ for some index set T of size 2^n, every function $f : \{0,1\}^n \to \mathbb{R}$ has the unique expansion

$$f = \sum_{t\in T} \langle f, b_t\rangle b_t \ , \tag{2.4}$$

see, e.g., Artin [Art91] or Lang [Lan93]. A natural orthonormal basis of this space is given by the *(normalized) Dirac functions* $\delta_a : \{0,1\}^n \to \{0,1\}$, $a \in \{0,1\}^n$, defined by

$$\delta_a(x) = \begin{cases} D(a)^{-1/2} & \text{if } a = x, \\ 0 & \text{otherwise.} \end{cases}$$

We have $\langle f, \delta_a \rangle_D = \sqrt{D(a)} f(a)$. Thus, $\|\delta_a\|_D = 1$ and for $a \neq b$,

$$\langle \delta_a, \delta_b \rangle_D = \sqrt{D(b)} \delta_a(b) = 0 \ .$$

However, there is another basis that is more important to us. In case that D is the uniform distribution on $\{0,1\}^n$, this is the so-called *Hadamard basis* or *Walsh basis* or *Rademacher basis* $(\chi_I \mid I \subseteq [n])$, defined by

$$\chi_I(x) = (-1)^{\sum_{i \in I} x_i} \tag{2.5}$$

for $x \in \{0,1\}^n$, see for example Bernasconi [Ber98].

For arbitrary non-degenerate D, we obtain a corresponding orthonormal basis $(\chi_I^D \mid I \subseteq [n])$ by applying the Gram-Schmidt orthonormalization procedure (see, e.g., Artin [Art91] or Lang [Lan93]) to $(\chi_I \mid I \subseteq [n])$ (e.g., in order of growing $|I|$).

For the special case that D is a product distribution on $\{0,1\}^n$ with

$$\Pr[x_i = 1] = d_i \in (0,1) \ ,$$

let $\sigma_i = \sqrt{d_i(1-d_i)}$ be the standard deviation of x_i. Bahadur [Bah61] has shown that $\chi_i^D : \{0,1\}^n \to \mathbb{R}$ is then given by

$$\chi_i^D(x) = \frac{d_i - x_i}{\sigma_i} = \begin{cases} \sqrt{\frac{d_i}{1-d_i}} & \text{if } x_i = 0, \\ -\sqrt{\frac{1-d_i}{d_i}} & \text{if } x_i = 1, \end{cases} \tag{2.6}$$

and for $I \subseteq [n]$, $\chi_I^D : \{0,1\}^n \to \mathbb{R}$ is given by $\chi_I^D = \prod_{i \in I} \chi_i^D$, see also Furst, Jackson, and Smith [FJS91]. By independence of the variables under product distributions, for all nonempty $I \subseteq [n]$, we have $\mathbb{E}[\chi_I^D] = \prod_{i \in I} \mathbb{E}[\chi_i^D] = 0$ (whereas $\chi_\emptyset^D \equiv 1$ and hence $\mathbb{E}[\chi_\emptyset^D] = 1$). It is useful to compute how $\chi_i^D(x)$ is affected by flipping the i^{th} bit: from (2.6), we obtain

$$\chi_i^D(x \oplus e_i) = -\frac{1-d_i}{d_i} \cdot \chi_i^D(x) \quad \text{for } x \in \{0,1\}^n \text{ with } x_i = 0 \ . \tag{2.7}$$

Indeed, $\chi_I^{U_n} = \chi_I$ for all $I \subseteq [n]$, and for the special case $d_i = 1/2$, (2.7) is immediate from (2.5).

The *Fourier transform* $\mathcal{F}_D : \mathbb{R}^{\{0,1\}^n} \to \mathbb{R}^{\mathcal{P}([n])}$ maps real-valued functions on the hypercube to real-valued functions on the power set of $[n]$ by

$$\mathcal{F}_D(f)(I) = \langle f, \chi_I^D \rangle_D$$

for $f : \{0,1\}^n \to \mathbb{R}$ and $I \subseteq [n]$. Although $\{0,1\}^n$ and $\mathcal{P}([n])$ are essentially the same, distinguishing them in this context is helpful to emphasize that the function f and its Fourier transform $\hat{f}$ are in a certain sense "of different types". $\mathcal{F}_D(f)(I)$ is called the *Fourier coefficient of f at I with respect to D*. In case that D is clear from the context, we also write

$$\hat{f}(I) = \mathcal{F}_D(f)(I) \; .$$

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

$$f(x) = \sum_{I \subseteq [n]} \hat{f}(I) \cdot \chi_I^D(x) \tag{2.8}$$

for all $x \in \{0,1\}^n$ is a special case of (2.4). Since the χ_I^D form a basis, the coefficients $\hat{f}(I)$ are unique in the following sense: whenever $f = \sum_{I \subseteq [n]} \alpha_I \chi_I$ for $\alpha_I \in \mathbb{R}$ (or even $\mathbb{C}$), then $\alpha_I = \hat{f}(I)$ for all $I \subseteq [n]$.

By orthonormality, for all $I, J \subseteq [n]$,

$$\widehat{\chi_I^D}(J) = \begin{cases} 1 & \text{if } I = J, \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore,

$$\mathrm{Var}[\chi_I^D] = \mathbb{E}[(\chi_I^D)^2] - \mathbb{E}[\chi_I^D]^2 = \langle \chi_I^D, \chi_I^D \rangle_D - 0 = 1 \; .$$

For $p \geq 1$, define the *p-norm* of f by

$$\|f\|_p = \mathbb{E}_{x \sim D}[|f(x)|^p]^{1/p} = \left(\sum_{x \in \{0,1\}^n} D(x) |f(x)|^p \right)^{1/p} \; . \tag{2.9}$$

We are mainly interested in the cases $p = 1$ and $p = 2$. In particular, $\|f\|_1$ is simply the average value of $|f(x)|$. Moreover, for $f, g : \{0,1\}^n \to \{0,1\}$, we have

$$\Pr_{x \sim D}[f(x) \neq g(x)] = \|f - g\|_1 = \|f - g\|_2^2 \; ,$$

whereas for $f, g : \{0,1\}^n \to \{-1,+1\}$, we have

$$\Pr_{x \sim D}[f(x) \neq g(x)] = \tfrac{1}{2}\|f - g\|_1 = \tfrac{1}{4}\|f - g\|_2^2 \; .$$

In both cases, for $\epsilon \geq 0$, we say that f is *ϵ-close* to g (with respect to D) if $\Pr_{x \sim D}[f(x) \neq g(x)] \leq \epsilon$ and *ϵ-far* from g otherwise. An intensively used feature of Fourier coefficients is *Parseval's equation*

$$\sum_{I \subseteq [n]} \hat{f}(I)^2 = \|f\|_2^2 = \sum_{x \in \{0,1\}^n} D(x) f(x)^2 \; . \tag{2.10}$$

Note that $\|f\|_D$ and $\|f\|_2$ coincide. In case that $f : \{0,1\}^n \to \{0,1\}$,

$$\|f\|_D^2 = \Pr[f = 1] = \hat{f}(\emptyset)$$

is the *bias* of f, whereas for $f : \{0,1\}^n \to \{-1,+1\}$, $\|f\|_D = 1$.

Fourier coefficients may also be interpreted in terms of statistical measures. Specifically, for two functions $f, g : \{0,1\}^n \to \mathbb{R}$, define the *covariance* of f and g with respect to D by

$$\mathrm{Cov}[f,g] = \mathbb{E}_{x\sim D}[(f(x) - \mathbb{E}[f]) \cdot (g(x) - \mathbb{E}[g])] = \mathbb{E}[f \cdot g] - \mathbb{E}[f] \cdot \mathbb{E}[g] \; .$$

Since $\mathbb{E}[\chi_I^D] = 0$ for all nonempty $I \subseteq [n]$, we obtain

$$\mathrm{Cov}[f, \chi_I^D] = \mathbb{E}[f \cdot \chi_I^D] = \langle f, \chi_I^D \rangle_D = \hat{f}(I) \; ,$$

i.e., $\hat{f}(I)$ is equal to the covariance of f and χ_I and thus measures how much f is *correlated* with χ_I^D. In fact, the *correlation coefficient* of f and g is defined by

$$\rho_{f,g} = \frac{\mathrm{Cov}[f,g]}{\sqrt{\mathrm{Var}[f] \cdot \mathrm{Var}[g]}} \; .$$

Thus, $\hat{f}(I) = \sqrt{\mathrm{Var}[f]} \cdot \rho_{f,\chi_I^D}$.

Alternatively, it is sometimes useful to think of the Fourier expansion (2.8) as a *polynomial representation* in the following sense. Let $z_1, \ldots, z_n$ be real variables. Depending on the values of $x_1, \ldots, x_n \in \{0,1\}$, we assign

$$z_i = (d_i - x_i)/\sigma_i$$

for $i \in [n]$. Clearly, $\chi_I^D(x) = \prod_{i\in I} z_i$. Hence, we may interpret χ_I as a real multilinear monomial in the variables z_i. By the uniqueness of the Fourier expansion, the representation of f as a multilinear real polynomial is also unique.

The following notions are used exclusively in Chapter 4, in which all concepts are assumed to map to $\Omega = \{0,1\}$.

Definition 2.2.1 (Fourier support). Let $f : \{0,1\}^n \to \{0,1\}$. The *Fourier support* of f is $\mathrm{supp}(\hat{f}) = \{I \subseteq [n] \mid \hat{f}(I) \neq 0\}$. The *Fourier support graph* $\mathrm{FSG}(f)$ of f is the subgraph of the n-dimensional Hamming cube induced by the sets in $\mathrm{supp}(\hat{f})$.

The following concept will be used extensively in Chapter 4. It also appears in the context of calculating a representation of a concept as a polynomial over the two-element field GF(2), see, e.g., Mossel et al. [MOS04].

Definition 2.2.2 (Expanded variable space). We define the *expanded variable space* of variables x_I, $I \subseteq [n]$, the values of which are determined by the values of the variables $x_1, \ldots, x_n$ in $\{0,1\}$ via

$$x_I = \bigoplus_{i \in I} x_i \; .$$

Thereby, we identify $x_{\{i\}}$ and x_i. Note that $\chi_I(x) = (-1)^{x_I}$.

Under the uniform distribution, instead of writing a function f in terms of the variables $x_1, \ldots, x_n$, one can also write it as a function of the variables x_I, $I \in \operatorname{supp}(\hat{f})$:

Lemma 2.2.3. *Let $D = U_n$, $f : \{0,1\}^n \to \mathbb{R}$ and $\mathcal{T} = \operatorname{supp}(\hat{f})$. Then there exists a function $g : \{0,1\}^{\mathcal{T}} \to \mathbb{R}$ such that for all $x \in \{0,1\}^n$,*

$$f(x_1, \ldots, x_n) = g(x_I \mid I \in \mathcal{T}) \; .$$

Proof. Define g by

$$g(x_I \mid I \in \mathcal{T}) = \sum_{I \in \operatorname{supp}(\hat{f})} \hat{f}(I) \cdot (-1)^{x_I} \; .$$

By (2.8), $f(x_1, \ldots, x_n) = g(x_I \mid I \in \mathcal{T})$ for all $x \in \{0,1\}^n$. □

The following lemma reveals a connection between vanishing Fourier coefficients of functions and their restrictions.

Lemma 2.2.4. *Let $D : \{0,1\}^n \to [0,1]$ be a non-degenerate product distribution, $f : \{0,1\}^n \to \{0,1\}$, $R \subseteq [n]$, $a \in \{0,1\}^R$, and $I \subseteq [n] \setminus R$. Then*

$$\widehat{f_a}(I) = \sum_{J \subseteq R} \hat{f}(I \cup J) \chi_J^D(a) \; .$$

Proof. By the Fourier expansion (2.8),

$$\begin{aligned} f_a &= \Big(\sum_{J \subseteq [n]} \hat{f}(J) \chi_J^D\Big)_a = \sum_{J \subseteq [n]} \hat{f}(J) \left(\chi_J^D\right)_a \\ &= \sum_{J_1 \subseteq [n] \setminus R} \sum_{J_2 \subseteq R} \hat{f}(J_1 \cup J_2) \left(\chi_{J_1 \cup J_2}^D\right)_a \\ &= \sum_{J_1 \subseteq [n] \setminus R} \sum_{J_2 \subseteq R} \hat{f}(J_1 \cup J_2) \chi_{J_2}^D(a) \chi_{J_1}^D(a) \; . \end{aligned}$$

By the uniqueness of the Fourier expansion (2.8), the claim follows (consider $J_1 = I$). □

Corollary 2.2.5. *Let $f : \{0,1\}^n \to \{0,1\}$, $R \subseteq [n]$, and $I^* \subseteq [n] \setminus R$. If for all $I \subseteq R$, $\hat{f}(I^* \cup I) = 0$, then for all $a \in \{0,1\}^R$, $\widehat{f_a}(I^*) = 0$.*

Proof. By Lemma 2.2.4, $\widehat{f_a}(I^*) = \sum_{I \subseteq R} \hat{f}(I^* \cup I) \chi_I(a) = 0$. □

2.3 Juntas

A function that depends only on a small subset of its variables is called a *junta.* For such functions, the function value is determined by the values of a small number of *relevant variables.* In this sense, juntas are a generalization of *dictatorship functions.* These are functions that depend only on a single variable. Dictatorship functions play an important role in *social choice theory.* We use juntas to model the presence of much irrelevant information in the learning data.

Definition 2.3.1 (Dependence/Relevance). A function $f : \{0,1\}^n \to \Omega$ *depends* on variable x_i (equivalently, x_i is *relevant* to f) if the restrictions $f_{x_i=0}$ and $f_{x_i=1}$ with x_i set to 0 and 1, respectively, are not equal. This is the case if and only if there exists an $x \in \{0,1\}^n$ with $f(x) \neq f(x \oplus e_i)$. A variable that is not relevant to f is called *irrelevant.* We denote the set of relevant variables of f by $\mathrm{rel}(f)$ and the set of irrelevant variables of f by $\mathrm{irrel}(f)$.

If f is clear from the context, we call a variable that is relevant to f simply *relevant.* The *base function* f' of f is the unique restriction of f to its relevant variables:

Definition 2.3.2 (Base function). Let $f : \{0,1\}^n \to \Omega$ and $a \in \{0,1\}^{[n]\setminus\mathrm{rel}(f)}$ be an arbitrary assignment (e.g., the constant zero assignment). Then the *base function* $f' : \{0,1\}^{\mathrm{rel}(f)} \to \Omega$ is given by $f' = f_a$.

Definition 2.3.3 (Junta). 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 $\mathcal{J}_d^n$.

A parity function f with $|\mathrm{rel}(f)| \leq d$ is called a *parity d-junta* or a *d-parity.* The class of parity d-juntas defined on n variables is denoted by PAR_d^n.

The class of monotone d-juntas is denoted by MON_d^n, and the class of juntas with symmetric base function is denoted by SYM_d^n.

The results of this section will be used for concepts $f : \{0,1\}^n \to \Omega$ with $\Omega = \{0,1\}$ or $\Omega = \{-1,+1\}$.

The next lemma has implicitly been proved by Mossel et al. [MOS04] for the uniform distribution. We state a more general form for non-degenerate product distributions. The lemma can be used to decide whether a variable is relevant to a function or not. Intuitively, it 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 2.3.4. *Let $f : \{0,1\}^n \to \Omega$ and $D : \{0,1\}^n \to [0,1]$ be a non-degenerate product distribution. 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 $\hat{f}(I) \neq 0$.*

Proof. We prove the contrapositions of the claim. Let $i \in [n]$. Since D is a product distribution, $\chi_I^D = \prod_{i \in I} \chi_i^D$ for all $I \subseteq [n]$ (see Section 2.2). If for all $I \subseteq [n]$ with $i \in I$, $\hat{f}(I) = 0$, then for all $x \in \{0,1\}^{[n]\setminus\{i\}}$,

$$\begin{aligned} f_{x_i=0}(x) &= \sum_{I \subseteq [n]} \hat{f}(I)(\chi_I^D)_{x_i=0}(x) \;=\; \sum_{I \subseteq [n]\setminus\{i\}} \hat{f}(I)(\chi_I^D)_{x_i=0}(x) \\ &= \sum_{I \subseteq [n]\setminus\{i\}} \hat{f}(I)(\chi_I^D)_{x_i=1}(x) \;=\; f_{x_i=1}(x) \; . \end{aligned}$$

Consequently, $f_{x_i=0} = f_{x_i=1}$, i.e., $x_i \notin \mathrm{rel}(f)$.

On the other hand, if $x_i \notin \mathrm{rel}(f)$, then $f_{x_i=0} = f_{x_i=1}$. Let $I \subseteq [n]$ with $i \in I$. Then, by (2.7), for $x \in \{0,1\}^n$ with $x_i = 0$, $\chi_I(x \oplus e_i) = -\frac{1-d_i}{d_i} \cdot \chi_I(x)$. Thus, we obtain

$$\begin{aligned} \hat{f}(I) &= \sum_{x \in \{0,1\}^n : x_i = 0} \big(D(x)f(x)\chi_I^D(x) + D(x \oplus e_i)f(x \oplus e_i)\chi_I^D(x \oplus e_i)\big) \\ &= \sum_{x' \in \{0,1\}^{[n]\setminus\{i\}}} \big(D(x') \cdot (1-d_i) \cdot f_{x_i=0}(x') \cdot \chi_I^D(x') \\ &\qquad\qquad -D(x') \cdot d_i \cdot f_{x_i=1}(x') \cdot \frac{1-d_i}{d_i} \cdot \chi_I^D(x')\big) \\ &= 0 \; . \end{aligned}$$

□

The Fourier coefficients of a function remain the same if one restricts the function to its relevant variables:

Lemma 2.3.5. *Let $f : \{0,1\}^n \to \Omega$ and $I \subseteq [n]$. If $I \not\subseteq \mathrm{rel}(f)$, then $\hat{f}(I) = 0$. If $I \subseteq \mathrm{rel}(f)$, then $\hat{f}(I) = \widehat{f'}(I)$.*

Proof. The case $I \not\subseteq \mathrm{rel}(f)$ follows from Lemma 2.3.4. Thus, it remains to consider $I \subseteq \mathrm{rel}(f)$. Recall that $f' = f_a$ for $a \in \{0,1\}^{\mathrm{irrel}(f)}$, $a \equiv 0$. By Lemma 2.2.4,

$$\widehat{f'}(I) = \widehat{f_a}(I) = \sum_{J \subseteq \mathrm{irrel}(f)} \hat{f}(I \cup J)\chi_J(a) \; .$$

Since for $J \subseteq \mathrm{irrel}(f)$ with $J \neq \emptyset$, $\hat{f}(I \cup J) = 0$, we have

$$\widehat{f'}(I) = \hat{f}(I)\chi_\emptyset(a) = \hat{f}(I) \; .$$

□

If D is the uniform distribution on $\{0,1\}^n$, then nonzero Fourier coefficients of d-juntas are of absolute size at least 2^{-d}:

Lemma 2.3.6. *Let $D = U_n$, $f : \{0,1\}^n \to \{0,1\}$, and $I \subseteq [n]$ such that $\hat{f}(I) \neq 0$. Then $|\hat{f}(I)| \geq 2^{-|\operatorname{rel}(f)|}$.*

Proof. By Lemma 2.3.5, $I \subseteq \operatorname{rel}(f)$ and $\hat{f}(I) = \widehat{f'}(I) = 2^{-d} \sum_{J \subseteq \operatorname{rel}(f)} f'(x)\chi_I(x)$, which is an integer multiple of 2^{-d}. □

Trivially, if $f, g : \{0,1\}^n \to \mathbb{R}$ are two d-juntas, then $f - g$ is a $2d$-junta. Thus, if $f - g$ differs from zero in some point x, then it does so in at least 2^{n-2d} further points which are obtained from x by arbitrarily setting the variables in $\operatorname{irrel}(f - g)$. Consequently, it is easy to show that $\Pr_{x \sim U_n}[f(x) \neq g(x)] \geq 2^{-2d}$. Interestingly, two distinct d-juntas cannot even differ in less than a 2^{-d} fraction of the inputs:

Lemma 2.3.7. *Let $f, g : \{0,1\}^n \to \{0,1\}$ be d-juntas. If $f \neq g$, then there are at least 2^{n-d} points $x \in \{0,1\}^n$ such that $f(x) \neq g(x)$ and thus,*

$$\Pr_{x \sim U_n}[f(x) \neq g(x)] \geq 2^{-d} .$$

Proof. We start by proving the claim for the case that $\operatorname{rel}(f) \cap \operatorname{rel}(g) = \emptyset$. In that case, if $g \equiv b$, $b \in \{0,1\}$, then $f \not\equiv b$ since $f \neq g$. This implies that there is an $a \in \{0,1\}^{\operatorname{rel}(f)}$ such that $f(a) = 1 - b$. Then for all assignments $a' \in \{0,1\}^{[n]\setminus\operatorname{rel}(f)}$, $f(a \sqcup a') = 1 - b \neq b = g(a \sqcup a')$. These are $2^{n-|\operatorname{rel}(f)|} \geq 2^{n-d}$ assignments.

Now assume that g is non-constant. Let $a_0, a_1 \in \{0,1\}^{\operatorname{rel}(g)}$ such that $g'(a_0) = 0$ and $g'(a_1) = 1$. Then for each $a \in \{0,1\}^{\operatorname{rel}(f)}$,

$$g'(a_{1-f'(a)}) = 1 - f'(a) \neq f'(a) .$$

Thus, for all $x \in \{0,1\}^{[n]\setminus(\operatorname{rel}(f)\cup\operatorname{rel}(g))}$ and all $a \in \{0,1\}^{\operatorname{rel}(f)}$,

$$f(a \sqcup a_{1-f'(a)} \sqcup x) = f'(a) \neq g'(a_{1-f'(a)}) = g(a \sqcup a_{1-f'(a)} \sqcup x) .$$

These are $2^{|\operatorname{rel}(f)|} \cdot 2^{n-|\operatorname{rel}(f)|-|\operatorname{rel}(g)|} = 2^{n-|\operatorname{rel}(g)|} \geq 2^{n-d}$ assignments.

Next, let us consider the case that $\operatorname{rel}(f)$ and $\operatorname{rel}(g)$ have nonempty intersection. Let $R = \operatorname{rel}(f) \cap \operatorname{rel}(g)$. Let $a \in \{0,1\}^n$ such that $f(a) \neq g(a)$ and let $a' = a|_R$. Then $f_{a'} \neq g_{a'}$. Let $r = \max\{|\operatorname{rel}(f)| - |R|, |\operatorname{rel}(g)| - |R|\}$. Clearly, $|\operatorname{rel}(f_{a'})| \leq r$, $|\operatorname{rel}(g_{a'})| \leq r$, and $\operatorname{rel}(f_{a'}) \cap \operatorname{rel}(g_{a'}) = \emptyset$. By the above considerations, $f_{a'}$ and $g_{a'}$ differ in at least

$$2^{n-|R|-r} = 2^{n-\max\{|\operatorname{rel}(f)|,|\operatorname{rel}(g)|\}} \geq 2^{n-d}$$

points, and so do f and g. □

2.4 Lowness and Fourier-accessibility

By Lemma 2.3.4, 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 nonempty subsets of size at most $d = |\operatorname{rel}(f)|$. However, there are $V(n,d) \approx n^d$ such subsets, which is $\Theta(n^d)$ for constant d, an amount that we would like to reduce. This leads us to the following definition:

Definition 2.4.1 ($\boldsymbol{\tau}$-lowness). Let $f : \{0,1\}^n \to \Omega$, $x_i \in \operatorname{rel}(f)$, and $\tau \in [n]$. 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 $x_i \in \operatorname{rel}(f)$ are τ-low for f. The set of τ-low d-juntas is denoted by $\mathcal{J}_d^n(\tau)$.

For τ-low concepts, one can find all relevant variables by investigating the Fourier coefficients up to level τ.

The special case of 1-lowness of a variable x_i has several equivalent characterizations. One of them is based on the *information gain* of x_i. This is a measure for the appropriateness of selecting x_i as a *split variable* in a greedy tree learning procedure to build a small decision tree for the target concept, see Rosell et al. [RHRP05] and references therein. Precisely, the *information gain* of x_i with respect to the distribution D is

$$I_D(f|x_i) = H_D(f) - H_D(f|x_i) \;,$$

where

$$H_D(f) = -\Pr[f=1] \cdot \log \Pr[f=1] - \Pr[f=0] \cdot \log \Pr[f=0]$$

(with all probabilities with respect to $x \sim D$) is the *binary entropy* of f (with respect to D) and

$$H_D(f|x_i) = \Pr[x_i = 0] \cdot H_D(f_{x_i=0}) + \Pr[x_i = 1] \cdot H_D(f_{x_i=1})$$

is the *binary entropy* of f *conditional on* x_i (with respect to D).

Lemma 2.4.2. *Let $D : \{0,1\}^n \to [0,1]$ be a non-degenerate product distribution and $f : \{0,1\}^n \to \{0,1\}$. Then the following statements are equivalent.*

(a) x_i is 1-low for f.

(b) $\hat{f}(x_i) \neq 0$.

(c) $\Pr[f_{x_i=0} = 1] \neq \Pr[f_{x_i=1} = 1]$, i.e., $f_{x_i=0}$ and $f_{x_i=1}$ have different bias.

(d) $\Pr[f_{x_i=0} = 1] \neq \Pr[f = 1]$ *or* $\Pr[f_{x_i=1} = 1] \neq \Pr[f = 1]$*, i.e., setting* x_i *to* $a \in \{0,1\}$ *changes the bias of* f *at least for* $a = 0$ *or for* $a = 1$*.*

(e) $I_D(f|x_i) > 0$*.*

Proof. (a) ⇔ (b) holds by definition.
(b) ⇔ (c): We have

$$\begin{aligned} \hat{f}(i) &= \sum_{x\in\{0,1\}^n} D(x)f(x)\chi_i^D(x) \\ &= \Pr[f(x) = 1 \wedge x_i = 0] \cdot \sqrt{d_i/(1-d_i)} \\ &\quad - \Pr[f(x) = 1 \wedge x_i = 1] \cdot \sqrt{(1-d_i)/d_i} \\ &= \sigma_i(\Pr[f(x) = 1 \mid x_i = 0] - \Pr[f(x) = 1 \mid x_i = 1]) \\ &= \sigma_i(\Pr[f_{x_i=0} = 1] - \Pr[f_{x_i=1} = 1]) \ . \end{aligned}$$

Thus, $\hat{f}(i) \neq 0$ implies $\Pr[f_{x_i=0} = 1] \neq \Pr[f_{x_i=1} = 1]$. Since D is assumed to be non-degenerate, $\sigma_i \neq 0$. Hence, also the converse holds.
(c) ⇔ (d): (c) ⇒ (d) is obvious. The other direction is proved by means of

$$\Pr[f = 1] = d_i \cdot \Pr[f_{x_i=1} = 1] + (1-d_i) \cdot \Pr[f_{x_i=0} = 1] \ :$$

If $\Pr[f_{x_i=0} = 1] = \Pr[f_{x_i=1} = 1]$, then $\Pr[f = 1] = \Pr[f_{x_i=0} = 1] = \Pr[f_{x_i=1} = 1]$.
(e) ⇔ (c): The equivalence of parts (e) and (c) has been proved by Rosell et al. [RHRP05, Lemma 2.1]. □

Condition (c) in Lemma 2.4.2 has been called *unbalanced with respect to* x_i by Fukagawa and Akutsu [FA05].

Next we show how to extend the equivalence of (a) and (d) in Lemma 2.4.2 to τ-lowness, providing a criterion for (inclusion-) minimal sets $I \neq \emptyset$ in the Fourier support $\mathrm{supp}(\hat{f})$.

Lemma 2.4.3. *Let* $f : \{0,1\}^n \to \{0,1\}$ *and* $I \subseteq [n]$ *such that for all* $J \subsetneq I$ *with* $J \neq \emptyset$*,* $\hat{f}(J) = 0$*. Then* $\hat{f}(I) \neq 0$ *if and only if there exists an assignment* $a \in \{0,1\}^I$ *such that* $\Pr[f_a = 1] \neq \Pr[f = 1]$*.*

Proof. By Lemma 2.2.4,

$$\Pr[f_a = 1] = \hat{f_a}(\emptyset) = \sum_{J\subseteq I} \hat{f}(J)\chi_J(a) = \hat{f}(\emptyset) + \hat{f}(I)\chi_I(a) \ .$$

On the other hand, $\Pr[f = 1] = \hat{f}(\emptyset)$. The claim follows. □

For a given concept class $\mathcal{C}$, finding the smallest τ such that $\mathcal{C} \subseteq \mathcal{J}_d^n(\tau)$ has attracted considerable interest in the past. The following example lists some known results.

Example 2.4.4. In this example, we assume that $D = U_n$ is the uniform distribution on $\{0,1\}^n$.

(a) Monotone functions are 1-low (a generalization of this statement to non-degenerate product distributions is proved in Lemma 5.6.2 on page 93). In symbols, this means that $\mathrm{MON}_d^n \subseteq \mathcal{J}_d^n(1)$. Even more: all concepts that are locally monotone are 1-low; these are functions that can be turned into a monotone function by negating some input variables. This includes all monomials and clauses of arbitrary literals. For monotone functions, the Fourier coefficient $\hat{f}(i)$ is equal to the *influence* of x_i as defined for example by Kahn, Kalai, and Linial [KKL88].

(b) 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 [BL97] and Mossel et al. [MOS04].

(c) At the other end, we have $\mathcal{J}_d^n(d) = \mathcal{J}_d^n$ (by Lemma 2.3.4), the class of all d-juntas on n variables.

(d) The parity function $f : \{0,1\}^n \to \{0,1\}$ defined by $f(x) = \bigoplus_{i \in I} x_i$ for some $I \subseteq [n]$ is $|I|$-low, but not $(|I|-1)$-low (by orthonormality of the functions χ_I, $I \subseteq [n]$).

(e) The class of all unbalanced d-juntas is contained in $\mathcal{J}_d^n((2/3) \cdot d)$ (see Mossel et al. [MOS04]).

(f) The class $\mathcal{S} = \mathrm{SYM}_d^n \setminus \{\chi_I \mid |I| \leq d\}$ of symmetric d-juntas that are not parity functions was shown to be contained in $\mathcal{J}_d^n((2/3) \cdot d)$ by Mossel et al. [MOS04]. This was improved by Lipton et al. [LMMV05] to $\mathcal{S} \subseteq \mathcal{J}_d^n((3/31) \cdot d)$. Very recently, Kolountzakis et al. [KMM05] showed that $\mathcal{S} \subseteq \mathcal{J}_d^n(O(d/\log d))$. Concerning lower bounds, even the question whether $\mathcal{S} \subseteq \mathcal{J}_d^n(O(1))$ is still open.

(g) By Lemma 2.2.3, every function $f : \{0,1\}^n \to \mathbb{R}$ can be written as a function of variables x_I, $I \in \mathrm{supp}(\hat{f})$. If f is *not* τ-low, then there exists a variable x_i such that $i \in I \in \mathrm{supp}(\hat{f})$ only for sets I of size larger than τ. In this interpretation, the value of x_i is "masked" by at least τ other bits each time it occurs.

The notion of τ-lowness is independent of the representation of Boolean function values:

Lemma 2.4.5. *Let $f : \{0,1\}^n \to \{0,1\}$ and $g = \rho \circ f$ with $\rho : \{0,1\} \to \{-1,+1\}$ defined by $\rho(x) = (-1)^x = 1 - 2x$. Then for $I \subseteq [n]$ with $I \neq \emptyset$, $\hat{g}(I) = -2\hat{f}(I)$. In particular, $\hat{f}(I) = 0$ if and only if $\hat{g}(I) = 0$. Furthermore, $\hat{g}(\emptyset) = 1 - 2\hat{f}(\emptyset)$.*

Proof. For $I \subseteq [n]$ with $I \neq \emptyset$, we have

$$\begin{aligned} \hat{g}(I) &= \sum_{x \in \{0,1\}^n} D(x) g(x) \chi_I^D(x) \\ &= \sum_{x \in \{0,1\}^n} D(x)(1 - 2f(x)) \chi_I^D(x) \\ &= \langle 1, \chi_I^D \rangle_D - 2\hat{f}(I) \;=\; -2\hat{f}(I)\,, \end{aligned}$$

where the latter equation follows from $\langle 1, \chi_I^D \rangle_D = \langle \chi_\emptyset^D, \chi_I^D \rangle_D = 0$. □

As a consequence, Lemma 2.4.2 is also valid for $f : \{0,1\}^n \to \{-1,+1\}$ (replacing 0 with $+1$ and 1 with -1 in all occurrences).

For our characterization of the functions to which Greedy is applicable in Chapter 4, we introduce the concept of *Fourier-accessibility.*

Definition 2.4.6 (Fourier-accessible). Let $f : \{0,1\}^n \to \{0,1\}$ and $i \in [n]$. Variable x_i is *accessible* (with respect to f) if there exists a sequence

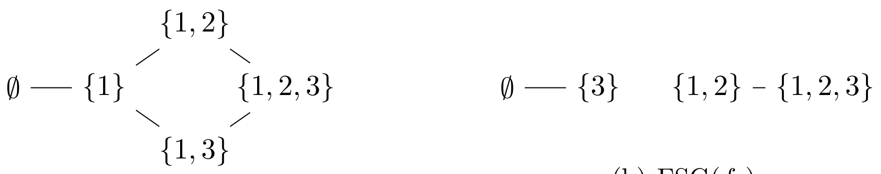
$$\emptyset = I_0 \subsetneq I_1 \subsetneq \ldots \subsetneq I_s \subseteq [n]$$

such that

1. $i \in I_s$,
2. for all $j \in [s]$, $|I_j \setminus I_{j-1}| = 1$, and
3. for all $j \in [s]$, $\hat{f}(I_j) \neq 0$.

The set of variables that are accessible with respect to f is denoted by $\mathrm{acc}(f)$, whereas the set of inaccessible variables with respect to f is denoted by $\mathrm{inacc}(f)$. The function f is *Fourier-accessible* if and only if every variable that is relevant to f is also accessible, i.e., $\mathrm{acc}(f) = \mathrm{rel}(f)$. In general, $\mathrm{acc}(f) \subseteq \mathrm{rel}(f)$.

If f is clear from the context, we call a variable that is accessible with respect to f simply *accessible.* The following lemma provides two conditions that are equivalent to Fourier-accessibility.



(a) FSG(f_1).

(b) FSG(f_2).

Figure 2.1: Fourier support graphs of functions f_1 and f_2 presented in Table 2.1.

Lemma 2.4.7. *Let $f : \{0,1\}^n \to \{0,1\}$ and $i \in [n]$. Then the following statements are equivalent:*

(a) f is Fourier-accessible.

(b) For all $x_i \in \operatorname{rel}(f)$, there exists an $I \subseteq [n]$ with $i \in I$ such that there is a path in the Fourier support graph $\operatorname{FSG}(f)$ from $\emptyset$ to I.

(c) The union of all subsets $I \in \operatorname{supp}(\hat{f})$ that belong to the connected component of $\emptyset$ in $\operatorname{FSG}(f)$ is equal to $\operatorname{rel}(f)$.

Proof. (a) ⇒ (b): Let $x_i \in \operatorname{rel}(f)$ and $\emptyset = I_0 \subsetneq I_1 \subsetneq \ldots \subsetneq I_s \subseteq [n]$ satisfy conditions 1.–3. in Definition 2.4.6. Let $I = I_s$. Then the sequence $\emptyset = I_0, \ldots, I_s = I$ is a path in $\operatorname{FSG}(f)$, provided that $\emptyset \in \operatorname{FSG}(f)$. But since

$$\hat{f}(\emptyset) = \Pr[f(x) = 1] \ ,$$

$\emptyset \in \operatorname{supp}(\hat{f})$ whenever $f \not\equiv 0$ (here it is essential that $\Omega = \{0,1\}$). In case that $f \equiv 0$, however, $\operatorname{rel}(f) = \emptyset$.
(b) ⇒ (c): Denote the union of all subsets $I \in \operatorname{supp}(\hat{f})$ that belong to the connected component of $\emptyset$ in $\operatorname{FSG}(f)$ by U. As a consequence of Lemma 2.3.4, $I \in \operatorname{supp}(\hat{f})$ implies that $I \subseteq \operatorname{rel}(f)$. Thus, $U \subseteq \operatorname{rel}(f)$. For the reverse inclusion, let $x_i \in \operatorname{rel}(f)$. By assumption, there exists $I \subseteq [n]$ with $i \in I$ and a path in $\operatorname{FSG}(f)$ from $\emptyset$ to I. Consequently, I is in the connected component of $\emptyset$ and hence $i \in I \subseteq U$.
(c) ⇒ (a): Let $x_i \in \operatorname{rel}(f)$. By assumption, there exists $I \in \operatorname{supp}(\hat{f})$ in the connected component of $\emptyset$ such that $i \in I$. Let $\emptyset = I_0, I_1, \ldots, I_s = I$ be a path in $\operatorname{FSG}(f)$. Then conditions 1.–3. in Definition 2.4.6 are satisfied. □

Example 2.4.8. (a) Simple examples of a Fourier-accessible function f_1 and a non-Fourier-accessible function f_2 are given in Table 2.1. The corresponding Fourier support graphs are presented in Figure 2.1. Note that f_2 is 2-Fourier-accessible.

Table 2.1: Examples of Boolean functions and their Fourier spectra.

	$f_1(x_1, x_2, x_3)$ $= x_1 \oplus (x_2 \wedge x_3)$	$f_2(x_1, x_2, x_3)$ $= (x_1 \oplus x_2) \wedge x_3$
$\hat{f}(\emptyset)$	$1/2$	$1/4$
$\hat{f}(1)$	$-1/4$	0
$\hat{f}(2)$	0	0
$\hat{f}(3)$	0	$-1/4$
$\hat{f}(\{1,2\})$	$-1/4$	$-1/4$
$\hat{f}(\{1,3\})$	$-1/4$	0
$\hat{f}(\{2,3\})$	0	0
$\hat{f}(\{1,2,3\})$	$1/4$	$1/4$

(b) If a function f is 1-low, then it is also Fourier-accessible. On the other hand, the function f_1 in Table 2.1 is Fourier-accessible, but not 1-low.

(c) A symmetric function f is 1-low if and only if it is Fourier-accessible. This is because for symmetric functions f, $\hat{f}(\pi(I)) = \hat{f}(I)$ for all permutations $\pi : [n] \to [n]$ and all $I \subseteq [n]$.

(d) An example of a function that is 2-low but not Fourier-accessible (and thus not 1-low either) is the *not-all-equal* function NAE : $\{0,1\}^n \to \{0,1\}$, defined by $\text{NAE}(x) = 1$ if and only if there exist $i, j \in [n]$ such that $x_i \neq x_j$. To see this, note that setting any variable x_i to 0 or 1 does obviously not change the bias. By Lemma 2.4.2, no variable x_i is 1-low for NAE, and thus no variable is accessible either. On the other hand, setting some variable x_i to 0 and another variable x_j to 1 turns the function into the constant 1-function and thus changes the bias. By Lemma 2.4.3, $\hat{f}(\{i,j\}) \neq 0$.

We weaken the notion of Fourier-accessibility by allowing the sets I_j in Definition 2.4.6 to grow by up to τ elements for some parameter τ:

Definition 2.4.9 (τ-Fourier-accessible). Let $f : \{0,1\}^n \to \{0,1\}$, $i \in [n]$, and $\tau \in [n]$. Variable x_i is *τ-accessible* (with respect to f) if there exists a sequence $\emptyset \subsetneq I_1 \subsetneq \ldots \subsetneq I_s \subseteq [n]$ such that

1. $i \in I_s$,

2. for all $j \in [s]$, $1 \leq |I_j \setminus I_{j-1}| \leq \tau$, and

3. for all $j \in [s]$, $\hat{f}(I_j) \neq 0$.

The variables that are not τ-accessible are called *τ-inaccessible*. The set of variables that are τ-accessible with respect to f is denoted by $\tau\text{-}\operatorname{acc}(f)$, whereas the set of variables that are τ-inaccessible with respect to f is denoted by $\tau\text{-}\operatorname{inacc}(f)$. The function f is *τ-Fourier-accessible* if and only if every variable that is relevant to f is also τ-accessible, i.e., $\tau\text{-}\operatorname{acc}(f) = \operatorname{rel}(f)$. In general, $\tau\text{-}\operatorname{acc}(f) \subseteq \operatorname{rel}(f)$.

CHAPTER 3

Learning from Noisy Data

This chapter is organized as follows. In Sections 3.1 and 3.2, we define the learning and noise models that are investigated in this thesis and introduce some standard tools from statistics. In Section 3.3, we show how one can approximate Fourier coefficients from noisy data. The noise operator and related concepts are introduced in Section 3.4. Subsequently, in Section 3.5, we present known upper and lower bounds on the sample size that is in general necessary to learn d-juntas from noise-free samples. Upper and lower bounds for learning from noisy data are provided in Section 3.6. That section also contains our generalization of a result due to Bshouty et al. [BJT03] concerning the applicability of an LMN-style learning algorithm [LMN93]. Our general characterization of learnability from noisy data, based on Bshouty et al.'s *noisy distance*, is presented in Section 3.7.

3.1 Noise-free Samples and Statistics

For the following definitions, let $f : \{0,1\}^n \to \Omega$ ($\Omega = \{0,1\}$ or $\Omega = \{-1,+1\}$) be a *target concept* and $D : \{0,1\}^n \to [0,1]$ be a probability distribution, the so-called *attribute distribution.* We start by defining *noise-free samples*:

Definition 3.1.1 (D-distributed sample). An *example* is a pair $(x, y) \in \{0,1\}^n \times \Omega$. The Boolean vector x is called the *attribute vector* and y is called the *classification* or *label* of the example. A sequence of examples $(x^k, y^k)_{k\in[m]}$ of m examples is called a *sample* of size m. If $x \sim D$ and $y = f(x)$, then the pair (x, y) is called a *(noise-free) D-distributed example for f*. A sequence S of m independent D-distributed examples for f is called a *(noise-free) D-distributed* *sample for f of size m.* U_n-distributed examples and samples are also called *uniformly distributed.*

If a sample S is explicitly given, we denote it in the following form:

$$S = \begin{pmatrix} x_1^1 & \dots & x_n^1 & \mid & y^1 \\ \vdots & \ddots & \vdots & & \vdots \\ x_1^m & \dots & x_n^m & \mid & y^m \end{pmatrix} . \tag{3.1}$$

Specifically, we use superscripts to indicate example indices and subscripts to indicate variable indices. Each row k consists of a single example (x^k, y^k), which in turn consists of n attribute values $x_1^k, \dots, x_n^k$ and a classification y^k.

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 [AS92]:

Lemma 3.1.2 (Hoeffding bound [Hoe63]). *Let X_i, $i \in [n]$, be mutually independent random variables taking values in the real interval $[a, b]$, $a < b$. Then for any $\epsilon \in [0, 1]$,*

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

Definition 3.1.3 (Empirical expectation). Let $t : \{0,1\}^n \times \Omega \to [-1, +1]$ and

$$S = (x^k, y^k)_{k \in [m]} \in (\{0,1\}^n \times \Omega)^m$$

be a sample. Define the *empirical expectation* of t given S as

$$\tilde{\mathbb{E}}_S[t] = \frac{1}{m} \sum_{k=1}^{m} t(x^k, y^k) .$$

Corollary 3.1.4. *Let $t : \{0,1\}^n \times \Omega \to [-1, +1]$ and $\delta, \epsilon > 0$. Let*

$$S = (x^k, y^k)_{k \in [m]} \in (\{0,1\}^n \times \Omega)^m$$

be a sample with examples generated according to a probability distribution

$$R : \{0,1\}^n \times \Omega \to [0, 1] .$$

If $m \geq 2 \cdot \ln(2/\delta) \cdot (1/\epsilon^2)$, then

$$\Pr\left(\left|\tilde{\mathbb{E}}_S[t] - \mathbb{E}_{(x,y)\sim R}[t(x, y)]\right| \geq \epsilon\right) \leq \delta .$$

This corollary is used for example when converting a statistical query algorithm into a PAC learning algorithm [Kea98]: the expectation of any $[-1,+1]$-valued function assigned to labeled examples can be approximated efficiently by relative frequencies from a small amount of examples. Our main application of this corollary is the approximation of Fourier coefficients.

Definition 3.1.5 (Empirical Fourier coefficient). Given a sample

$$S = (x^k, y^k)_{k\in[m]} \in (\{0,1\}^n \times \Omega)^m ,$$

define the *empirical Fourier coefficient of f at I given S* by

$$\tilde{f}_S(I) = \frac{1}{m}\sum_{k=1}^{m} \chi_I(x^k) \cdot y^k . \tag{3.2}$$

Lemma 3.1.6. *Let $I \subseteq [n]$ and S be a uniformly distributed sample of size $m \geq 2 \cdot \ln(2/\delta) \cdot (1/\epsilon^2)$. Then*

$$|\tilde{f}_S(I) - \hat{f}(I)| \leq \epsilon$$

with probability at least $1-\delta$.

Proof. Define $t : \{0,1\}^n \times \{-1,+1\} \to \{-1,+1\}$ by $t(x,y) = \chi_I(x) \cdot y$. The claim follows from Corollary 3.1.4. □

Definition 3.1.7 (Consistency). Let $S = (x^k, y^k)_{j\in[m]} \in (\{0,1\}^n \times \Omega)^m$ be a sample. A concept $h : \{0,1\}^n \to \Omega$ is *consistent* with S if for all $k \in [m]$, $y^k = h(x^k)$.

The following lemma is due to Blumer, Ehrenfeucht, Haussler, and Warmuth [BEHW87] and has now become a standard fact in algorithmic learning theory. It serves as a basis for the principle of *Occam's razor.*

Lemma 3.1.8 (Blumer et al. [BEHW87]). *Let $\mathcal{C}$ be a class of concepts $f : \{0,1\}^n \to \Omega$. Let $f \in \mathcal{C}$, $D : X \to [0,1]$ be a probability distribution on X, and $\delta, \epsilon > 0$. Let S be a D-distributed sample for f of size*

$$m \geq \frac{1}{\epsilon} \ln \frac{|\mathcal{C}|}{\delta} \tag{3.3}$$

and $h : X \to Y$ be any concept that is consistent with S. Then

$$\Pr_{x\sim D}[h(x) \neq f(x)] < \epsilon$$

with probability at least $1-\delta$ (taken over the set of D-distributed samples of size m for f).

Subsequent to the appearance of Blumer et al.'s work on the principle of Occam's razor [BEHW87], the same authors published one of the most influential papers in algorithmic learning theory [BEHW89], in which they essentially replace $\ln|\mathcal{C}|$ in (3.3) by the Vapnik-Chervonenkis dimension of $\mathcal{C}$. This yields an asymptotically tight bound on the number of examples necessary to learn. For our purposes, however, it suffices to work with Lemma 3.1.8.

3.2 Learning and Noise Models

Additionally to the target concept f and the attribute distribution D, we fix an *attribute noise distribution* $P : \{0,1\}^n \to [0,1]$ and a *classification noise rate* $\eta \in [0,1]$.

Definition 3.2.1 (($\boldsymbol{P,\eta}$)-noisy sample). Let $x \sim D$, $\xi \sim P$, and $\zeta \in \Omega$ with $\zeta \sim \eta$. In case that $\Omega = \{0,1\}$, the pair $(x \oplus \xi, f(x) \oplus \zeta)$ is called a *D-distributed (P,η)-noisy example* for f, whereas in case that $\Omega = \{-1,+1\}$, the pair $(x \oplus \xi, f(x) \cdot \zeta)$ 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 $\xi \sim P$ to the attribute vector x (componentwisely modulo 2) and flipping the classification y according to the classification noise bit $\zeta \sim \eta$.

A *$(P,0)$-noisy example* is corrupted only by attribute noise but not by classification noise. Examples that suffer from classification noise with rate η but not from attribute noise are denoted *$(-,\eta)$-noisy.* In these cases, we speak of *pure attribute noise* and *pure classification noise*, respectively. 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 $\eta = 0$. Hence, noise-free examples are $(-,0)$-noisy.

Definition 3.2.2 (Learning algorithm). Let $\delta \in (0,1]$ and $\epsilon \in [0,1]$, called the *confidence parameter* and the *accuracy parameter*, respectively. An algorithm $\mathcal{A}$ *learns* the class $\mathcal{C}$ with confidence $1-\delta$ and accuracy $1-\epsilon$ 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 $h : \{0,1\}^n \to \Omega$ 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.,

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

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 [Val84]. Learning with accuracy $\epsilon = 0$ 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.

A learning algorithm is said to be *attribute-efficient* if its sample size is polynomial in $\log n$, $2^{|\operatorname{rel}(f)|}$, $1/\delta$, and $1/\epsilon$. It has *polynomial running time* if its running time is polynomial in n, $2^{|\operatorname{rel}(f)|}$, $1/\delta$, and $1/\epsilon$. In all applications, the dependence on δ is in fact only $O(\log(1/\delta))$.

Definition 3.2.3 (Learnability of a concept class). A concept class $\mathcal{C}$ is *learnable* (with confidence $1-\delta$ and accuracy $1-\epsilon$ 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 $1-\epsilon$ from D-distributed (P,η)-noisy samples of size m in time t). It is *exactly learnable* if it is learnable with accuracy 1.

In the remainder of this subsection, we take a closer look at noise distributions and their properties. Bshouty et al. [BJT03, Lemma 1] have observed that attribute and classification noise may be reduced to "attribute- and concept-dependent" classification noise:

Lemma 3.2.4 ([BJT03]). *Let $P : \{0,1\}^n \to [0,1]$ be an attribute noise distribution and $\eta \in [0,1]$ be a classification noise rate. Let $x \sim U_n$, $\xi \sim P$, and $\zeta \sim \eta$ be independent random variables. Then the variables $(x \oplus \xi, f(x) \cdot \zeta)$ and $(x, f(x \oplus \xi) \cdot \zeta)$ are identically distributed.*

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 [GS95]. 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

$$P(\xi_1, \ldots, \xi_n) = \prod_{i:\xi_i=1} p_i \cdot \prod_{i:\xi_i=0} (1-p_i) = \prod_{i=1}^{n} p_i^{\xi_i} \cdot (1-p_i)^{1-\xi_i} \, .$$

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.

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 always assume that $\eta \neq 1/2$. The case $\eta > 1/2$ can be reduced to $\eta' < 1/2$ by negating all classifications and then multiplying with noise bits ζ drawn according to $\eta' = 1 - \eta$. Nevertheless, it has to be known a priori whether $\eta < 1/2$ or $\eta > 1/2$.

The knowledge of our learning algorithms about the noise distribution P and the noise rate η varies from exact distributions to quite weak assumptions such as $|1 - 2\eta| \geq \gamma_b$ for some $\gamma_b > 0$. Specifically, the greedy algorithm studied in Chapter 4 does not need any knowledge about the noise parameters at all. However, the sample size needed to learn successfully always has to depend on these parameters, as lower bounds show (see Theorem 3.6.1 and Theorem 6.4.2).

We often need to estimate, for a set $I \subseteq [n]$ of attribute indices, how likely an odd number of bits in I is flipped when attribute noise is applied. Therefore we introduce

$$p_I = \Pr_{\xi \sim P}[\chi_I(\xi) = -1] = \Pr_{\xi \sim P}\left[\bigoplus\nolimits_{i \in I} \xi_i = 1\right] . \tag{3.4}$$

Moreover, we denote $p_{\{i\}}$ by p_i, which is consistent with the case of product random attribute noise. If P is a product distribution such that all p_i differ from $1/2$, then also p_I differs from $1/2$:

Lemma 3.2.5. *Let $P : \{0,1\} \to [0,1]$ be a product distribution with rates $p_1, \ldots, p_n$ and $I \subseteq [n]$. Then*

$$1 - 2p_I = \prod_{i \in I}(1 - 2p_i) . \tag{3.5}$$

Moreover, if there exists $\gamma > 0$ such that for all $i \in [n]$, $|1 - 2p_i| \geq \gamma$, then $|1 - 2p_I| \geq \gamma^{|I|}$ for all $I \subseteq [n]$.

Proof. For $|I| \leq 1$, the claim is true. For $|I| \geq 2$, let $i \in I$ and $I' = I \setminus \{i\}$. Then

$$\begin{aligned}
1 - 2p_I &= 1 - 2\Pr_{\xi \sim P}[\chi_I(\xi) = -1] \\
&= 1 - 2\Pr_{\xi \sim P}[\chi_{I'}(\xi) = 1] \cdot \Pr_{\xi \sim P}[\xi_i = -1] \\
&\qquad -2\Pr_{\xi \sim P}[\chi_{I'}(\xi) = -1] \cdot \Pr_{\xi \sim P}[\xi_i = 1] \\
&= 1 - 2(1 - p_{I'})p_i - 2p_{I'}(1 - p_i) \;=\; 4p_{I'}p_i - 2p_{I'} - 2p_i + 1 \\
&= (1 - 2p_i)(1 - 2p_{I'}) .
\end{aligned}$$

By induction hypothesis, $1 - 2p_{I'} = \prod_{i \in I'}(1 - 2p_i)$, and the claim follows. □

Inspired by the result of the previous lemma, we define γ_a-bounded probability distributions:

Definition 3.2.6 (γ_a-bounded probability distribution). Let $\gamma_a \in (0,1]$. A probability distribution $P : \{0,1\}^n \to [0,1]$ is γ_a*-bounded* if for all $I \subseteq [n]$,

$$|1 - 2p_I| \geq \gamma_a^{|I|} \ .$$

Thus, if P is a product distribution with rates $p_1, \ldots, p_n$ that satisfy the condition $|1 - 2p_i| \geq \gamma_a > 0$ for all $i \in [n]$, then P is γ_a-bounded. The following example shows that Lemma 3.2.5 is not valid for *arbitrary* distributions (where we define $p_i = \Pr_{\xi \sim P}[\xi_i = 1]$).

Example 3.2.7. 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 $p_2 = P(01) + P(11) = 1/4$. Consequently, $(1 - 2p_1)(1 - 2p_2) = (1/2) \cdot (1/2) = 1/4 \neq 0$. However,

$$p_{\{1,2\}} = \Pr_{\xi \sim P}[\xi_1 \oplus \xi_2 = 1] = P(01) + P(10) = 1/2 \ ,$$

i.e., $1 - 2p_{\{1,2\}} = 0$. If P is an attribute noise distribution, this 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 more recoverable from any number of P-noisy attribute vectors. We will revert to this issue in Section 3.7.

On the other hand, also non-product distributions P may of course be γ_a-bounded:

Example 3.2.8. Let $n = 2$ and $P : \{0,1\}^2 \to [0,1]$ be defined by

$$P(00) = 1/2, \ P(01) = 1/8, \ P(10) = 1/4, \ \text{and } P(11) = 1/8 \ .$$

Then

$$\begin{aligned} p_1 &= P(10) + P(11) = 3/8 \ , \\ p_2 &= P(01) + P(11) = 1/4 \ , \text{ and} \\ p_{\{1,2\}} &= P(01) + P(10) = 3/8 \ . \end{aligned}$$

Consequently, $|1 - 2p_1| = 1/4$, $|1 - 2p_2| = 1/2$, and $|1 - 2p_{\{1,2\}}| = 1/4$ (note that always $p_\emptyset = 0$ and hence $|1 - 2p_\emptyset| = 1$). Thus, P is 1/4-bounded. Furthermore, P is not a product distribution since $P(11) = 1/8$, but $p_1 \cdot p_2 = 3/32$.

In addition to the probability p_I that a parity function χ_I applied to a noise vector takes the value -1, we will also often refer to the expected value of $\chi_I(\xi)$ when $\xi \sim P$: let

$$\lambda_I = \mathbb{E}_{\xi\sim P}[\chi_I(\xi)] = 1 - 2p_I. \tag{3.6}$$

We have

$$\lambda_I = \mathbb{E}_{\xi\sim P}[\chi_I(\xi)] = 2^{-n} \sum_{\xi\in\{0,1\}^n} 2^n P(\xi)\chi_I(\xi) = 2^n \hat{P}(I)\ , \tag{3.7}$$

where the Fourier transform is taken with respect to the uniform distribution.

Interestingly, we can conclude from (3.7) that the converse of Lemma 3.2.5 is also true. Although we do not use this result any further, we include it here as a nice application of Fourier analysis in probability theory.

Lemma 3.2.9. *Let $P : \{0,1\}^n \to [0,1]$ be a probability distribution that satisfies (3.5) for all $I \subseteq [n]$. Then P is a product distribution.*

Proof. By (2.8),

$$P = \sum_{I\subseteq[n]} \hat{P}(I)\chi_I = 2^{-n} \sum_{I\subseteq[n]} \lambda_I \chi_I = 2^{-n} \sum_{I\subseteq[n]} (1-2p_I)\chi_I\ .$$

Thus, the mapping $P \mapsto (p_I \mid I \subseteq [n])$ is injective with inverse function

$$(p_I \mid I \subseteq [n]) \mapsto 2^{-n} \sum_{I\subseteq[n]} (1-2p_I)\chi_I\ .$$

If (3.5) is valid for all $I \subseteq [n]$, then we can compute all p_I, $I \subseteq [n]$, from $p_1, \dots, p_n$. Let P' be the product distribution with rates $p'_i = p_i$ for all $i \in [n]$. By the preceding argument, $p'_I = p_I$ for all $I \subseteq [n]$. Since the mapping above is injective, $P = P'$. Hence, P is a product distribution. □

3.3 Approximation of Fourier Coefficients from Noisy Data

In Lemma 3.1.6, we have shown that the empirical Fourier coefficient $\tilde{f}_S(I)$ is a close approximation to $\hat{f}(I)$ if the noise-free sample S is sufficiently large. If the sample is (P,η)-noisy, then $\tilde{f}_S(I)$ approximates

$$\mathbb{E}_{x\sim U_n, \xi\sim P, \zeta\sim\eta}[\chi_I(x \oplus \xi) \cdot f(x) \cdot \zeta]\ .$$

Lemma 3.3.1 (Bshouty et al. [BJT03]). *For all $I \subseteq [n]$, we have*

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

Proof. We include a proof for the sake of completeness. By assumption, the attribute vector x, the attribute noise vector ξ, and the classification noise bit ζ are pairwise independent. Furthermore, $\chi_I(x \oplus \xi) = \chi_I(x) \cdot \chi_I(\xi)$. Thus,

$$\mathbb{E}_{x\sim U_n,\xi\sim P,\zeta\sim\eta}[\chi_I(x \oplus \xi) \cdot f(x) \cdot \zeta] = \mathbb{E}_{x\sim U_n}[\chi_I(x)f(x)] \cdot \mathbb{E}_{\xi\sim P}[\chi_I(\xi)] \cdot \mathbb{E}_{\zeta\sim\eta}[\zeta] \; .$$

It remains to compute each of the three factors: by definition,

$$\mathbb{E}_{x\sim U_n}[\chi_I(x)f(x)] = \hat{f}(I) \quad \text{and} \quad \mathbb{E}_{\xi\sim P}[\chi_I(\xi)] = 1 - 2p_I \; .$$

Finally, $\mathbb{E}_{\zeta\sim\eta}[\zeta] = (1 - \eta) \cdot 1 + \eta \cdot (-1) = 1 - 2\eta$. □

A straightforward application of the Hoeffding bound (see Lemma 3.1.2) yields

Lemma 3.3.2. *Let $I \subseteq [n]$ and $m \geq 2 \cdot \ln(2/\delta) \cdot (1/\epsilon^2)$. Then*

$$|\tilde{f}_S(I) - (1 - 2p_I)(1 - 2\eta)\hat{f}(I)| \leq \epsilon$$

with probability at least $1 - \delta$.

Proof. Define $t : \{0,1\}^n \times \{-1,+1\} \to \{-1,+1\}$ by $t(x,y) = \chi_I(x) \cdot y$. The claim follows from Corollary 3.1.4. □

Thus, we can infer $\hat{f}(I)$ by dividing $\tilde{f}(I)$ by $(1 - 2p_I)(1 - 2\eta)$. This is possible if and only if $p_I \neq 1/2$ and $\eta \neq 1/2$. Requesting η to be different from 1/2 is reasonable (even necessary) as we have discussed in Section 3.2. Unfortunately, it can happen that $p_I = 1/2$ for some I (even if $\Pr_{\xi\sim P}[\xi_i = 1] \neq 1/2$ for all $i \in [n]$, see Example 3.2.7), yielding a concept class $\mathcal{C}$ and an attribute noise distribution P such that $\mathcal{C}$ is (information-theoretically) not learnable from $(P, 0)$-noisy samples, as we will show in Theorem 3.7.3. In contrast, things look much nicer for product distributions P with noise rates p_i that are all different from 1/2, as we have seen in Lemma 3.2.5. In this setting, $p_I \neq 1/2$ for all $I \subseteq [n]$.

3.4 Noise Operator, Noisy Distance, and Noise Stability

Let us now introduce some mathematical tools that will be used to prove upper and lower sample bounds. We start with the *noise operator*, which allows us to view notions such as noisy examples, noisy distance, and noise stability in a unified framework.

Definition 3.4.1 (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

$$T_P(f)(x) = \mathbb{E}_{\xi \sim P}[f(x \oplus \xi)] \tag{3.8}$$

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 the attribute distribution D, then $T_P(f)(x)$ is the expected value of the classification of the corrupted attribute vector $x \oplus \xi$. 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 $p_1 = \ldots = p_n$, this operator has been extensively studied in the literature, e.g., by Kahn, Kalai, and Linial [KKL88], Benjamini et al. [BKS99], Mossel and O'Donnell [MO03], and O'Donnell [O'D03].

We show how the Fourier coefficients of $T_P(f)$ are related to those of f. Recall that $\lambda_I = \mathbb{E}_{\xi \sim P}[\chi_I(\xi)]$ for all $I \subseteq [n]$.

Lemma 3.4.2. *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

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

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

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

□

Using the Fourier expansion (2.8) and Parseval's equality (2.10), the following corollary is immediate:

Corollary 3.4.3. *Let $f : \{0,1\}^n \to [-1,1]$ and $P : \{0,1\}^n \to [0,1]$ be an attribute noise distribution. Then*

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

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

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

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

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

Proof. Part (a) follows by Fourier expansion (2.8) and Lemma 3.4.2 (b), part (b) is immediate from the definitions, and part (c) follows from the definitions and from Parseval's equality (2.10). The first inequality of part (d) 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 (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). □

Another corollary concerns the invertibility of T_P:

Corollary 3.4.4. *Let $P : \{0,1\}^n \to [0,1]$ be an attribute noise distribution.*

(a) The kernel $\ker(T_P) = \{f \in \mathbb{R}^{\{0,1\}^n} \mid T_P(f) = 0\}$ *of T_P is equal to the linear span of the parity functions χ_I with $\lambda_I = 0$:*

$$\ker(T_P) \;=\; \langle\, \chi_I \mid I \subseteq [n] : \lambda_I = 0 \,\rangle \,.$$

(b) T_P is invertible if and only if $\lambda_I \ne 0$ for all $I \subseteq [n]$.

Proof. (a) By Lemma 3.4.2 (a), for each $I \subseteq [n]$, we have $T_P(\chi_I) = \lambda_I \cdot \chi_I$. Consequently, $\langle\, \chi_I \mid I \subseteq [n] : \lambda_I = 0 \,\rangle \subseteq \ker(T_P)$. For the reverse inclusion, consider $f \in \ker(T_P)$, i.e., $T_P(f) = 0$. Then, by Lemma 3.4.2 (b),

$$\lambda_I \hat{f}(I) = \widehat{T_P(f)}(I) = 0$$

for all $I \subseteq [n]$. Consequently, if $\lambda_I \ne 0$, then $\hat{f}(I) = 0$. By Fourier expansion (2.8),

$$f = \sum_{I\subseteq[n]} \hat{f}(I)\chi_I = \sum_{I:\lambda_I\ne 0} \hat{f}(I)\chi_I \,.$$

Thus, f is contained in the linear span

$$\langle\, \chi_I \mid I \subseteq [n] : \lambda_I = 0 \,\rangle \ .$$

(b) This is immediate from part (a). □

Bshouty et al. [BJT03] have introduced a parameter $\Delta_P(f, g)$, which they have called the *noisy distance* between concepts f and g with respect to P. It turns out that $\Delta_P(f, g)$ is precisely half the 1-norm of the noise operator T_P applied to the difference of f and g:

$$\Delta_P(f, g) \ = \ \tfrac{1}{2}\, \mathbb{E}_{x\sim U_n}[\,|\, \mathbb{E}_{\xi\sim P}[f(x \oplus \xi) - g(x \oplus \xi)]\,|\ \ = \ \tfrac{1}{2}\, \|T_P(f - g)\|_1 \ . \quad (3.9)$$

We have proposed an interpretation of the expression $\mathbb{E}_{\xi\sim P}[f(x\oplus\xi)] = T_P(f)(x)$ in the beginning of this subsection. The value $\Delta_P(f, g)$ measures the expected difference between noisy examples for f and for g. The smaller this expected difference becomes, the harder it is to tell f and g apart on the basis of random noisy examples.

For any $\epsilon > 0$, Bshouty et al. have defined $\Delta_P^\epsilon(\mathcal{C})$ to be the minimum noisy distance between ϵ-far concepts inside $\mathcal{C}$:

$$\Delta_P^\epsilon(\mathcal{C}) \ = \ \min\,\{\Delta_P(f, g) \mid f, g \in \mathcal{C} : \Pr_{x\sim U_n}[f(x) \neq g(x)] > \epsilon\} \ . \quad (3.10)$$

In our notation this is equal to

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

i.e., $\Delta_P^\epsilon(\mathcal{C})$ measures how close ϵ-far concepts in $\mathcal{C}$ can become when T_P is applied to them. In addition, for $\alpha : 2^{[n]} \to [-1, 1]$, Bshouty et al. [BJT03] have defined the *α-attenuated power spectrum* of f by $s_\alpha(f) = \sum_{I\subseteq[n]} \alpha(I)^2 \hat{f}(I)^2$. For $\alpha(I) = \lambda_I$ as introduced in (3.6), they have defined $s_P(f) = s_\alpha(f)$, i.e.,

$$s_P(f) = \sum_{I\subseteq[n]} \lambda_I^2 \hat{f}(I)^2 \ , \quad (3.11)$$

and have proved a useful relationship between $s_\alpha(f - g)$ and the noisy distance $\Delta_P(f, g)$ (see Theorem 3.4.5). Again, it turns out that their definitions can be rephrased in terms of norms of the noise operator applied to the difference between concepts f and g: we simply have

$$s_P(f) = \|T_P(f)\|_2^2$$

by Corollary 3.4.3 (c).

In analogy to $\Delta_P^\epsilon(\mathcal{C})$, the minimum $s_P(f-g)$ of ϵ-far concepts in $\mathcal{C}$ is

$$S_P^\epsilon(\mathcal{C}) \;=\; \min\{s_P(f-g) \mid f,g \in \mathcal{C} : \Pr_{x\sim U_n}[f(x) \neq g(x)] > \epsilon\} \; ,$$

or, equivalently,

$$S_P^\epsilon(\mathcal{C}) \;= \min\left\{\|T_P(f-g)\|_2^2 \;\mid\; f,g \in \mathcal{C} : \tfrac{1}{2}\|f-g\|_1 > \epsilon\right\} .$$

Theorem 3.4.5 ([BJT03]). *Let $P : \{0,1\}^n \to [0,1]$ be a probability distribution.*

(a) For any functions $f,g : \{0,1\}^n \to \{-1,+1\}$,

$$\frac{1}{4}\, s_P(f-g) \leq \Delta_P(f,g) \leq \frac{1}{2}\,\sqrt{s_P(f-g)}\; .$$

(b) For every concept class $\mathcal{C}$ and every $\epsilon > 0$,

$$\tfrac{1}{4}\, S_P^\epsilon(\mathcal{C}) \leq \Delta_P^\epsilon(\mathcal{C}) \leq \tfrac{1}{2}\,\sqrt{S_P^\epsilon(\mathcal{C})}\; .$$

Although this theorem has been proved by Bshouty et al. [BJT03], we think that it is worthwhile to rephrase its proof in terms of the noise operator T_P and its norms since this adds a little more structure. With the prerequisites we have established, the proof appears more natural than the way it has been presented by Bshouty et al.

Proof. We deduce part (a) from Corollary 3.4.3 (d): from $|(f(x)-g(x))/2| \leq 1$ for all $x \in \{0,1\}^n$, it follows that

$$\|T_P((f-g)/2)\|_2^2 \leq \|T_P((f-g)/2)\|_1 \leq \|T_P((f-g)/2)\|_2\; .$$

The claim follows from

$$\begin{aligned}\|T_P((f-g)/2)\|_2^2 &= s_P(f-g)/4\; ,\\ \|T_P((f-g)/2)\|_1 &= \Delta_P(f,g)\; , \text{ and}\\ \|T_P((f-g)/2)\|_2 &= \sqrt{s_P(f)}/2\; .\end{aligned}$$

Part (b) is now a simple consequence of the definitions of $\Delta_P(\mathcal{C})$ and $S_P(\mathcal{C})$. □

We conclude this subsection by introducing the notion of noise stability with respect to arbitrary attribute noise distributions. For $x \in \{0,1\}^n$,

$$\Pr_{\xi\sim P}[f(x \oplus \xi) \neq f(x)]$$

is the probability that the classification of x is flipped by the attribute noise process. Since

$$\begin{aligned} T_P(f)(x) &= \sum_{\xi\in\{0,1\}^n} P(\xi) f(x\oplus\xi) \\ &= \sum_{\xi: f(x\oplus\xi)=f(x)} P(\xi) f(x) - \sum_{\xi: f(x\oplus\xi)=-f(x)} P(\xi) f(x) \\ &= f(x)\cdot\big(\Pr_{\xi\sim P}[f(x\oplus\xi)=f(x)] - \Pr_{\xi\sim P}[f(x\oplus\xi)\neq f(x)]\big) \\ &= f(x)\cdot\big(1-2\Pr_{\xi\sim P}[f(x\oplus\xi)\neq f(x)]\big)\ , \end{aligned}$$

we obtain that

$$\Pr_{\xi\sim P}[f(x\oplus\xi)\neq f(x)] = (1-|T_P(f)(x)|)/2\ .$$

The value $|1-2\Pr_{\xi\sim P}[f(x\oplus\xi)\neq f(x)]| = |T_P(f)(x)|$ measures how far the noise process is from turning the classification $f(x)$ into a random bit that does not reveal any information about $f(x)$. The *noise stability* is the minimum of this measure, taken over all $x\in\{0,1\}^n$:

Definition 3.4.6 (Noise stability). Let $\mathcal{C}$ be a concept class, $f\in\mathcal{C}$, and P be an attribute noise distribution. Define the *noise stability* $\Gamma_P(f)$ of f with respect to P by

$$\Gamma_P(f) \;=\; \min\left\{ \left|1-2\Pr_{\xi\sim P}[f(x\oplus\xi)\neq f(x)]\right| \;\Big|\; x\in\{0,1\}^n \right\}$$

and $\Gamma_P(\mathcal{C}) = \min\{\Gamma_P(f) \mid f\in\mathcal{C}\}$.

Alternative characterizations are given by

$$\begin{aligned} \Gamma_P(f) &= \min\left\{ \left|\mathbb{E}_{\xi\sim P}[f(x\oplus\xi)\cdot f(x)]\right| \;\Big|\; x\in\{0,1\}^n \right\} \\ &= \min\left\{ |T_P(f)(x)| \;\Big|\; x\in\{0,1\}^n \right\}\ . \end{aligned} \tag{3.12}$$

The related notion of noise sensitivity has been studied by Kahn et al. [KKL88], Benjamini et al. [BKS99], Mossel and O'Donnell [MO03], and O'Donnell [O'D03]. For product distributions P with equal rates $p_1 = \ldots = p_n = \epsilon \in [0,1]$, O'Donnell [O'D03] has defined

$$\mathrm{NS}_\epsilon(f) = \Pr_{x\sim U_n,\xi\sim P}[f(x\oplus\xi)\neq f(x)]\ .$$

Thus, $\Gamma_P(f) = |1-2\mathrm{NS}_\epsilon(f)|$ in this case. In general, $\Gamma_P(f)$ is a measure for the noise stability under arbitrary noise distributions P. We will use this measure in Chapter 6.

3.5 Sample Bounds for Learning Juntas from Noise-free Data

Almuallim and Dietterich [AD94, Theorem 1] have proved that given a number of examples that is at least polynomial in $\log n$, $2^{|\operatorname{rel}(f)|}$, $\log(1/\delta)$, and $1/\epsilon$, any hypothesis consistent with the sample is ϵ-close to the target concept with probability at least $1-\delta$. Their proof is based on the sample upper bound by Blumer et al. [BEHW87] (Lemma 3.1.8). For uniformly distributed examples, this means that with probability at least $1-\delta$, the target concept is the *only* concept consistent with the sample:

Lemma 3.5.1 ([AD94]). *Let $f : \{0,1\}^n \to \{0,1\}$, $d = |\operatorname{rel}(f)|$, and $\delta > 0$. Let S be a uniformly distributed sample of size*

$$m \geq 2^{d+1}\left(\ln\frac{1}{\delta} + d\ln n + 2^d \ln 2\right) \tag{3.13}$$

for f. Then with probability at least $1-\delta$, f is the only d-junta that is consistent with S. In particular, (3.13) is implied by

$$m \geq 2^{2d+1} \ln\frac{n}{\delta}\ .$$

As a lower bound for learning the relevant attributes from noise-free data, we cite a Theorem of Almuallim and Dietterich [AD94, Theorem 3], which is based on the computation of a lower bound on the Vapnik-Chervonenkis dimension of $\mathcal{J}_d^n$.

Theorem 3.5.2 ([AD94]). *To learn $\mathcal{J}_d^n$ distribution-freely with confidence $1-\delta$ and accuracy $1-\epsilon$, a sample size of*

$$\Omega\left(\tfrac{1}{\epsilon}(\ln\tfrac{1}{\delta} + \ln 2^d \ln n + 2^d)\right)$$

is necessary.

3.6 Upper and Lower Bounds for Learning from Noisy Data

Let us now come to infer bounds for the noisy learning scenario. The following lower bound is due to Bshouty et al. [BJT03, Theorem 2]:

Theorem 3.6.1 ([BJT03]). *Let $\mathcal{C}$ be a concept class, P be an attribute noise distribution and $\delta, \epsilon > 0$. Then any algorithm that learns the class $\mathcal{C}$ with confidence $1 - \delta$ and accuracy $1 - \epsilon$ from $(P, 0)$-noisy samples requires a sample complexity of*

$$\Omega\left(\frac{1-2\delta}{\Delta_P^{2\epsilon}(\mathcal{C})}\right) .$$

In the following, we describe an approach to obtain an upper bound for learning in fairly general situations. Let $\mathcal{T}_\epsilon \subseteq \mathcal{P}([n])$ such that

$$\sum_{I \in \mathcal{T}_\epsilon} \hat{f}(I)^2 \geq 1 - \epsilon$$

for all $f \in \mathcal{C}$. An *LMN-style algorithm* estimates, for each $I \in \mathcal{T}_\epsilon$, the Fourier coefficient $\hat{f}(I)$ (e.g., using empirical Fourier coefficients (3.2)), and then computes from these estimates a hypothesis h that is ϵ-close to f with high probability. This hypothesis is build via the Fourier expansion formula (2.8). See Bshouty et al. [BJT03] for a more precise definition of "LMN-style algorithm", and consult the seminal paper of Linial, Mansour, and Nisan [LMN93] in which such an algorithm has been proposed first.

In all applications we are aware of, $\mathcal{T}_\epsilon$ always consists of all sets up to a certain size depending on ϵ. In such a setting, the algorithm is also referred to as a *lowdegree algorithm* in the literature. However, one might construct concept classes $\mathcal{C}$ for which index sets $\mathcal{T}_\epsilon$ exist with $\sum_{I \in \mathcal{T}_\epsilon} \hat{f}(I)^2 \geq 1 - \epsilon$ but $\epsilon \mapsto \mathcal{T}_\epsilon$ is not efficiently computable, particularly in case of growing n. By saying that $\mathcal{C}$ is *LMN-style learnable with index set $\mathcal{T}_\epsilon$*, we exclude (by definition) these pathological cases. Bshouty et al. [BJT03, Theorem 8] have shown the following:

Theorem 3.6.2 ([BJT03]). *Let $\mathcal{C}$ be a concept class that is closed under complement and suppose that $\mathcal{C}$ is LMN-style learnable using index set $\mathcal{T}_\epsilon \subseteq [n]$. Then for every $\delta, \epsilon > 0$ such that $\{\chi_I \mid I \in \mathcal{T}_\epsilon\} \subseteq \mathcal{C}$, $\mathcal{C}$ is learnable with confidence $1 - \delta$ and accuracy $1 - 2\epsilon$ from uniformly distributed (P, η)-noisy samples in time polynomial in $|\mathcal{T}_\epsilon|$, $1/\Delta_P^\epsilon$, $\log(1/\delta)$, $1/\epsilon$, and $1/|1 - 2\eta|$.*

There are two minor caveats to the statement of Theorem 3.6.2: first, $\mathcal{C}$ is required to be closed under complement. Second, $\mathcal{C}$ is required to contain all parities indexed by $\mathcal{T}_\epsilon$. We remedy these shortcomings by replacing Δ_P^ϵ with

$$\lambda = \min\{|\lambda_I| \mid I \subseteq \mathcal{T}_\epsilon\} . \tag{3.14}$$

Note that under the assumptions of Theorem 3.6.2, $\Delta_P^\epsilon(\mathcal{C}) \leq \lambda$ since for all $I \subseteq [n]$,

$$\Delta_P(\chi_I, -\chi_I) = \|T_P(\chi_I)\|_1 = \|\lambda_I \chi_I\|_1 = |\lambda_I|$$

Algorithm 3.1 Noisy-LMN$_{\mathcal{T}}$

```
input $S = ((x_1^k, \ldots, x_n^k), y^k)_{k \in [m]}, P, \eta$
for $I \in \mathcal{T}$ do
    $\tilde{f}_S(I) \leftarrow \frac{1}{m} \sum_{k=1}^{m} \chi_I(x^k) \cdot y^k$
output hypothesis
```

$$\textsc{Noisy-LMN}_{\mathcal{T}}(x) = \operatorname{sgn} \sum_{I \in \mathcal{T}} (1 - 2p_I)^{-1} (1 - 2\eta)^{-1} \tilde{f}_S(I) \chi_I(x)$$

by (3.9) and Lemma 3.4.2 (a). The algorithm Noisy-LMN$_{\mathcal{T}}$ (Algorithm 3.1) will be used to derive upper bounds on the sample and time complexity of general learning bounds.

Theorem 3.6.3. *Let $\mathcal{C}$ be a class of concepts $\{0,1\}^n \to \{-1,+1\}$, $\epsilon > 0$, and $\mathcal{T}_\epsilon \subseteq \mathcal{P}([n])$ such that for each $f \in \mathcal{C}$, $\sum_{I \in \mathcal{T}_\epsilon} \hat{f}(I)^2 \geq 1 - \epsilon$. Then* Noisy-LMN$_{\mathcal{T}_\epsilon}$ *learns $\mathcal{C}$ with confidence $1 - \delta$ and accuracy $1 - 2\epsilon$ from uniformly distributed (P, η)-noisy samples using sample complexity and running time polynomial in $|\mathcal{T}_\epsilon|$, $1/\lambda$, $1/|1 - 2\eta|$, $\log(1/\delta)$, and $1/\epsilon$, with λ as defined in (3.14).*

Proof. The proof is basically the same as the proof of Theorem 3.6.2 presented by Bshouty et al. [BJT03], except for our improvements concerning the requirements for $\mathcal{C}$.

Let $\mathcal{T} = \mathcal{T}_\epsilon$ and $\rho = \lambda(1 - 2\eta)\sqrt{\epsilon/|\mathcal{T}|}$. For each $I \in \mathcal{T}$, let

$$\beta_I = (1 - 2p_I)^{-1}(1 - 2\eta)^{-1} \tilde{f}_S(I)$$

and define $h(x) = \sum_{I \in \mathcal{T}} \beta_I \chi_I(x)$ for $x \in \{0,1\}^n$. By Lemma 3.3.2, for each $I \in \mathcal{T}$, with probability at least $1 - \delta/|\mathcal{T}|$,

$$|\beta_I - \hat{f}(I)| \leq (1 - 2p_I)^{-1}(1 - 2\eta)^{-1}\rho \leq \sqrt{\epsilon/|\mathcal{T}|}\ ,$$

provided that

$$m \geq 2 \cdot \ln(2|\mathcal{T}|/\delta) \cdot (1/\rho^2) = 2 \cdot \ln(2|\mathcal{T}|/\delta) \cdot \lambda^{-2}(1 - 2\eta)^{-2}\epsilon^{-1} \cdot |\mathcal{T}|\ .$$

The latter amount of examples is polynomial in the parameters as claimed in the statement of the theorem. Hence, with probability at least $1 - \delta$,

$$|\beta_I - \hat{f}(I)| \leq \epsilon/|\mathcal{T}|$$

for all $I \in \mathcal{T}$ simultaneously. In this case, since $\mathrm{sgn}(h(x)) \neq f(x)$ implies $|h(x) - f(x)| \geq 1$, we obtain

$$\begin{aligned}
\Pr_{x\sim U_n}[\mathrm{sgn}(h(x)) \neq f(x)] &\leq 2^{-n} \sum_{x\in\{0,1\}^n} |h(x) - f(x)|^2 \\
&= \sum_{I\subseteq[n]} (\hat{h}(I) - \hat{f}(I))^2 \\
&= \sum_{I\in\mathcal{T}} (\beta_I - \hat{f}(I))^2 + \sum_{I\in\mathcal{P}([n])\setminus\mathcal{T}} \hat{f}(I)^2 \\
&\leq \sum_{I\in\mathcal{T}} \frac{\epsilon}{\mathcal{T}} + \epsilon \;=\; 2\epsilon\,,
\end{aligned}$$

where the first equality is Parseval's equation (2.10). □

Although sample and time complexity can be exponential in n, the method described will prove useful as part of our noise-tolerant learning algorithm for juntas (see Section 5.4).

3.7 A Characterization of Learnability from Noisy Data

We start this subsection by stating that if T_P maps two concepts f and g to the same function, then the pairs $(x \oplus \xi, f(x))$ and $(x \oplus \xi, g(x))$ are identically distributed:

Lemma 3.7.1 ([BJT03]). *Let $f, g : \{0,1\}^n \to \Omega$ such that $T_P(f - g) = 0$. Then for $x \sim U_n$ and $\xi \sim P$, the random pairs $(x \oplus \xi, f(x))$ and $(x \oplus \xi, g(x))$ are identically distributed.*

Proof. We provide a self-contained proof since the result has only been implicitly proved by Bshouty et al. [BJT03]. We first show that the pairs $(x, f(x \oplus \xi))$ and $(x, g(x \oplus \xi))$ are identically distributed. By definition of T_P, $T_P(f - g) = 0$ implies $\mathbb{E}_\xi[f(x \oplus \xi)] = \mathbb{E}_\xi[g(x \oplus \xi)]$ for all $x \in \{0,1\}^n$. In case that $\Omega = \{0,1\}$,

$$\begin{aligned}
\Pr_{\xi\sim P}[f(x \oplus \xi) = 1] &= \mathbb{E}_{\xi\sim P}[f(x \oplus \xi)] \\
&= \mathbb{E}_{\xi\sim P}[g(x \oplus \xi)] \\
&= \Pr_{\xi\sim P}[g(x \oplus \xi) = 1]
\end{aligned}$$

for all $x \in \{0,1\}^n$. Similarly, if $\Omega = \{-1,+1\}$, then

$$\begin{aligned}\Pr_{\xi\sim P}[f(x\oplus\xi) = -1] &= \tfrac{1}{2}(1 - \mathbb{E}_{\xi\sim P}[f(x\oplus\xi)]) \\ &= \tfrac{1}{2}(1 - \mathbb{E}_{\xi\sim P}[g(x\oplus\xi)]) \\ &= \Pr_{\xi\sim P}[g(x\oplus\xi) = -1]\end{aligned}$$

for all $x \in \{0,1\}^n$. Hence, for all $a \in \{0,1\}^n$ and $b \in \Omega$,

$$\begin{aligned}\Pr_{x\sim U_n,\xi\sim P}[x = a \wedge f(x\oplus\xi) = b] &= \Pr_{x\sim U_n}[x = a]\cdot\Pr_{x\sim U_n,\xi\sim P}[f(x\oplus\xi) = b \mid x = a] \\ &= 2^{-d}\Pr_{\xi\sim P}[f(a\oplus\xi) = b] \\ &= 2^{-d}\Pr_{\xi\sim P}[g(a\oplus\xi) = b] \\ &= \Pr_{x\sim U_n,\xi\sim P}[x = a \wedge g(x\oplus\xi) = b]\ .\end{aligned}$$

Consequently, $(x, f(x\oplus\xi))$ and $(x, g(x\oplus\xi))$ are identically distributed. By Lemma 3.2.4, also $(x\oplus\xi, f(x))$ and $(x\oplus\xi, g(x))$ are identically distributed. □

We tighten the lower and upper bounds of Bshouty et al. [BJT03] stated in Theorems 3.6.1 and 3.6.2 by providing the following characterization of learnability from uniformly distributed examples in the presence of attribute and classification noise:

Theorem 3.7.2. *Let $\mathcal{C}$ be a concept class, P be an attribute noise distribution, and $\eta \in [0,1]$ be a classification noise rate. Then $\mathcal{C}$ can be learned with accuracy $1-\epsilon$ from uniformly distributed (P,η)-noisy samples if and only if $\Delta_P^\epsilon(\mathcal{C}) > 0$ and $\eta \neq 1/2$.*

Proof. The "only if part" can be derived directly from the definition of Δ_P^ϵ: if this quantity vanishes, then there are distinct concepts $f, g \in \mathcal{C}$ such that $\Delta_P(f,g) = 0$. This implies that $\|T_P(f-g)\|_1 = 0$ and hence $T_P(f-g) = 0$. By Lemma 3.7.1, $(x\oplus\xi, f(x))$ and $(x\oplus\xi, g(x))$ with $x \sim U_n$ and $\xi \sim P$ are identically distributed, which means that also $(x\oplus\xi, f(x)\cdot\zeta)$ and $(x\oplus\xi, g(x)\cdot\zeta)$ (with $\Pr[\zeta = -1] = \eta$) are identically distributed. Hence, f and g are information-theoretically indistinguishable under attribute noise that is distributed according to P.

Now let $\Delta_P^\epsilon(\mathcal{C}) > 0$. If $\mathcal{C}$ is closed under complement and contains all parities, then Theorem 3.6.2 (with $\mathcal{T}_\epsilon = 2^{[n]}$) immediately yields the "if part" of the claim. However, all that is available to us for general classes $\mathcal{C}$ is Theorem 3.6.3, which we cannot apply since $\Delta_\epsilon(\mathcal{C}) > 0$ does not imply $\lambda = \min\{|\lambda_I| \mid I \subseteq [n]\} > 0$ (a counter-example is provided in Theorem 6.3.5). We thus have to take a different line since we cannot approximate any coefficients $\hat{f}(I)$ with $\lambda_I = 0$. Note that we are only interested in an information-theoretic result here; running time and sample complexity play no role.

The idea is now that it suffices to distinguish between any two concepts $f, g \in \mathcal{C}$ with $\|f - g\|_1 > 2\epsilon$. Let

$$\mathcal{T} = \{I \subseteq [n] \mid \lambda_I \neq 0\} \quad \text{and} \quad \lambda' = \min\{|\lambda_I| \mid I \in \mathcal{T}\} \ .$$

Let $f \in \mathcal{C}$ and $S = (x^k \oplus \xi^k, y^k \cdot \zeta^k)_{k \in [m]}$ be a uniformly distributed (P, η)-noisy sample for f of sufficiently large size m. We describe how to recover from S a hypothesis h that is ϵ-close to f with high probability: first, for all $I \in \mathcal{T}$, we calculate

$$\beta_I = \frac{\tilde{f}_S(I)}{|\lambda_I| \cdot |1 - 2\eta|} = (|\lambda_I| \cdot |1 - 2\eta| \cdot m)^{-1} \sum_{k=1}^{m} y^k \cdot \zeta^k \cdot f(x^k \oplus \xi^k) \ .$$

Then we check for each $h \in \mathcal{C}$ whether

$$\forall I \in \mathcal{T} : |\hat{h}(I) - \beta_I| \leq 2^{-n-1} \ .$$

If so, we output h. If we do not find such an h, then we output "failure".

By Lemma 3.3.2, if

$$m \geq 2 \cdot \ln(2^{n+1}/\delta) \cdot 2^{2n+4} (\lambda' \cdot |1 - 2\eta|)^{-2} \ ,$$

then with probability at least $1 - 2^{-n} \cdot \delta$,

$$|\beta_I - \hat{f}(I)| \leq 2^{-n-2} \ . \tag{3.15}$$

Consequently, with probability at least $1 - \delta$, (3.15) holds simultaneously for all $I \in \mathcal{T}$, which we assume in the following. In this case, f is among the candidates to be output. On the other hand, let $h \in \mathcal{C}$ with $\|f - h\|_1 > 2\epsilon$. Then $\|T_P(f - h)\|_1 \geq 2\Delta_P^\epsilon(\mathcal{C}) > 0$ and thus $T_P(f - h) \neq 0$. Since by Corollary 3.4.3,

$$T_P(f - h) = \sum_{I \subseteq [n]} \lambda_I (\hat{f}(I) - \hat{h}(I)) \chi_I \ ,$$

there has to be an $I \in \mathcal{T}$ such that $\hat{f}(I) \neq \hat{h}(I)$. Thus, $|\hat{f}(I) - \hat{h}(I)| \geq 2^{-n}$ and hence

$$|\hat{h}(I) - \beta_I| \geq |\hat{h}(I) - \hat{f}(I)| - |\hat{f}(I) - \beta_I| \geq 2^{-n} - 2^{-n-2} > 2^{-n-1} \ .$$

Consequently, h is not output. Overall, with probability at least $1 - \delta$, some hypothesis $h \in \mathcal{C}$ that is ϵ-close to f will be output. □

If $\Delta^{\epsilon}_{P}(\mathcal{C}) = 0$, then $\mathcal{C}$ cannot be learned from finitely many (P, η)-noisy examples with accuracy $\epsilon/2$ by Theorem 3.6.1, proving a slightly weaker result than the first direction of the previous theorem.

Finally, we prove that both possibilities $\Delta^{\epsilon}_{P}(\mathcal{C}) > 0$ and $\Delta^{\epsilon}_{P}(\mathcal{C}) = 0$ can occur for arbitrarily small $\epsilon > 0$. On the one hand, this proves that there are indeed classes that cannot be learned at all in the presence of noise. On the other hand, we show that γ_a-bounded distributions P (for $\gamma_a > 0$) have $\Delta^{\epsilon}_{P}(\mathcal{C}) > 0$ and thus in principle admit to learn arbitrary concept classes $\mathcal{C}$ from noisy examples.

Theorem 3.7.3. *There is a concept class $\mathcal{C}$ and an attribute noise distribution P such that $\mathcal{C}$ is not learnable from uniformly distributed $(P, 0)$-noisy samples. In addition, P may be chosen such that $p_i < 1/2$ for all $i \in [n]$.*

Proof. Consider the attribute noise distribution P given in Example 3.2.7. Let $f(x) = \chi_{\{1,2\}}(x) = (-1)^{x_1+x_2}$ and choose $\mathcal{C} = \{f, -f\}$. By Corollary 3.4.3 (c),

$$\|T_P(2f)\|_2^2 = 2 \sum_{I \subseteq \{1,2\}} \lambda_I^2 \widehat{\chi_{\{1,2\}}}(I)^2 = 2\lambda_{\{1,2\}}^2 = 1 - 2p_{\{1,2\}} = 0 ,$$

hence also $\|T_P(2f)\|_1 = 0$, i.e., $T_P(f - (-f)) = 0$. This implies that $(x \oplus \xi, f(x))$ and $(x \oplus \xi, -f(x))$ are identically distributed by Lemma 3.7.1. Hence, f and $-f$ are information-theoretically indistinguishable under P-attribute noise. □

On the other hand, if the probability for a certain bit to be flipped is 1/2, this does not automatically force $\Delta^{\epsilon}_{P}(\mathcal{C})$ to be zero: e.g., the variable under consideration may be irrelevant. As another example, consider the AND-function of a subset I of $[n]$ with x_1 being flipped with probability 1/2 and the rest not being changed at all. It is then easy to check (with sufficiently many examples) whether x_1 is relevant or not, although it is flipped with probability 1/2 (and thus is perfectly random): if the other variables do not admit a consistent AND-function, then x_1 must be relevant; otherwise, not.

Complementing Theorem 3.7.3, we show that any concept class is learnable under γ_a-bounded attribute noise and classification noise different from 1/2.

Theorem 3.7.4. *Let $\mathcal{C}$ be the class of all concepts $\{0,1\}^n \to \Omega$, $\Omega = \{0,1\}$ or $\Omega = \{-1,+1\}$. Let P be a γ_a-bounded attribute noise distribution and $\eta \neq 1/2$ be a classification noise rate. Then $\mathcal{C}$ is exactly learnable from uniformly distributed (P, η)-noisy samples.*

Proof. We have to show that $\Delta^{0}_{P}(\mathcal{C}) > 0$. Let $f, g \in \mathcal{C}$ with $f \neq g$. By assumption, $|\lambda_I| = |1 - 2p_I| \geq \gamma_a^{|I|} > 0$ for all $I \subseteq [n]$. Hence, by Corollary 3.4.4, T_P is invertible. Therefore, $T_P(f - g) \neq 0$. Consequently,

$$\Delta^{0}_{P}(\mathcal{C}) = \min \left\{ \|T_P(f - g)\| \;\middle|\; f, g \in \mathcal{C} \wedge f \neq g \right\} > 0 .$$

□

CHAPTER 4

The Greedy Method

The main result of this chapter is a concise characterization of the class of target concepts for which a simple greedy algorithm (called Greedy in this thesis) is able to infer the relevant variables. The idea behind the greedy algorithm is to reduce the learning problem to the combinatorial Set Cover problem and solve the latter by a well-known greedy algorithm. The characterization is based on a property of the Fourier spectrum of the target concept, which we call *Fourier-accessibility*, and which we have introduced in Definition 2.4.6. We present how to extend the algorithm to cope with the class of τ-Fourier-accessible concepts for $\tau \geq 1$. Furthermore, it is shown that Greedy is very robust against noise corrupting its input data.

This chapter is organized as follows. The reduction to Set Cover and the Greedy algorithm are presented in Section 4.1. Section 4.2 provides the major lemmas used in the proof of our main results for Greedy, which are presented in Section 4.3 for noise-free data. In Section 4.4, we extend the greedy approach to τ-Fourier accessible concepts. The scenario of noisy input data is treated in Section 4.5. Results for the static variant Greedy Ranking are presented in Section 4.6, followed by some comments about the applicability of the opposite strategy—Modest Ranking—in Section 4.7.

In this chapter, we are only concerned with finding the relevant attributes of the target concept. The task of constructing a hypothesis is deferred to Chapter 5. Furthermore, all concepts map to the range $\Omega = \{0, 1\}$. As a consequence, classification noise bits ζ will also be elements of $\{0, 1\}$, drawn with $\Pr[\zeta = 1] = \eta$ and $\Pr[\zeta = 0] = 1 - \eta$ for some classification noise rate $\eta \in [0, 1]$, which we indicate by $\zeta \sim \eta$. The classification y of a random example

$(x, y) \in \{0,1\}^n \times \{0,1\}$ is then affected by addition of ζ modulo 2 ("XOR-ed"). Thus, each classification is independently affected by noise with probability η.

4.1 Reduction to Set Cover and the Greedy Algorithm

In the following, we assume that all attributes and function values take binary values. The generalization of the definitions to larger domains and codomains is straightforward. Moreover, whenever randomness comes into play, we assume examples to be uniformly distributed.

Definition 4.1.1 (Functional relations graph). With a sample

$$S = (x^k, y^k)_{k\in[m]} \in (\{0,1\}^n \times \{0,1\})^m \ ,$$

we associate the *functional relations graph* $G_S = (V, E)$, which is defined as follows. Its vertices correspond to the examples of S, i.e., $V = [m]$. They are partitioned into the subset of examples $A^{(0)}$ with $y^k = 0$, and the examples $A^{(1)}$ with $y^k = 1$. G_S is the complete bipartite graph with the vertex set partition $[m] = A^{(0)} \cup A^{(1)}$, i.e.,

$$E = \left\{\{k, \ell\} \mid k, \ell \in [m], y^k \neq y^\ell\right\} \ .$$

Given S, our primary goal is to determine a set of variables $R \subseteq \{x_1, \dots, x_n\}$ of minimum size such that there exists *some* concept $g : \{0,1\}^n \to \{0,1\}$ with $\operatorname{rel}(g) \subseteq R$ that is consistent with the sample. In this case, R is said to *explain the sample.* Note that, in general, g need not be identical to the original concept f, nor need the set R contain all relevant variables of f.

In order to find an explaining set of variables, we have to specify, for each edge $\{k, \ell\} \in E$, a relevant variable that differs in x^k and x^ℓ. Such a variable is said to *explain the edge.* Formally, an edge $\{k, \ell\} \in E$ *can be covered* by an attribute x_i if and only if $x_i^k \neq x_i^\ell$. The set of edges that can be covered by x_i is denoted by E_i, i.e.,

$$E_i = \left\{\{k, \ell\} \in E \mid k, \ell \in [m], x_i^k \neq x_i^\ell\right\} \ .$$

The *characteristic vector* of an edge $e = \{k, \ell\} \in E$ is

$$c(e) = (c_1(e), \dots, c_n(e)) = \left(x_1^k \oplus x_1^\ell, \dots, x_n^k \oplus x_n^\ell\right) \ . \tag{4.1}$$

It is sometimes referred to as a *conflict* which may be *covered* by any variable x_i such that $c_i(e) = 1$, see, e.g., Almuallim and Dietterich [AD94].

A set R of variables thus explains the sample S if and only if these variables explain all edges. The previous discussion is summarized by the following lemma:

Table 4.1: A sample of size five.

k	x_1^k	x_2^k	x_3^k	x_4^k	x_5^k	x_6^k	$y^k = f(x^k)$
1	1	0	0	1	0	1	1
2	0	0	1	0	1	1	0
3	1	0	1	1	0	0	1
4	0	1	1	0	1	1	1
5	0	1	0	1	1	0	0

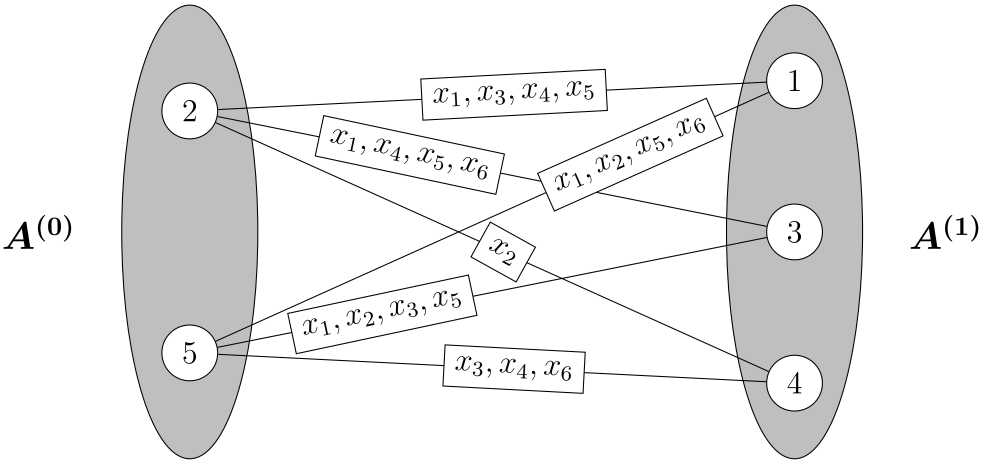


Figure 4.1: Functional relations graph of the sample given in Table 4.1. Each edge is labeled by the variables that can cover it.

Lemma 4.1.2. *Let $S \in (\{0,1\}^n \times \{0,1\})^m$ be a sample, $R \subseteq \{x_1, \ldots, x_n\}$, and E be the edge set of the functional relations graph G_S. Then R explains S if and only if $E = \cup_{x_i \in R} E_i$.*

Example 4.1.3. Consider the sample S given in Table 4.1. The corresponding functional relations graph G_S is presented in Figure 4.1. For instance, the edge $\{2,1\}$ may be covered by variables x_1, x_3, x_4, and x_5. Hence, its characteristic vector is

$$c(\{2,1\}) = (1,0,1,1,1,0) .$$

A set R that covers the edges of G_S is, e.g., $\{x_2, x_4\}$. Hence, $\{x_2, x_4\}$ explains the sample. Moreover, no other set of size at most two can cover all edges of G_S. In contrast, e.g., the set $R = \{x_1, x_3, x_4, x_5\}$ does not explain S since the edge $\{2,4\}$ is not covered by any of these variables.

Algorithm 4.1 Greedy

```
 1: input $S = ((x_1^k, \ldots, x_n^k), y^k)_{k \in [m]}$
 2: $E \leftarrow \{\{k, \ell\} \mid k, \ell \in [m], y^k \neq y^\ell\}$
 3: $R \leftarrow \emptyset$
 4: while $E \neq \emptyset$ do
 5:   for $i = 1$ to $n$ do
 6:     $E_i \leftarrow \{\{k, \ell\} \in E \mid x_i^k \neq x_i^\ell\}$
 7:   select $x_i$ with maximum $|E_i|$
 8:   $E \leftarrow E \setminus E_i$
 9:   $R \leftarrow R \cup \{x_i\}$
10: output Greedy$(S) = R$
```

Lemma 4.1.2 provides a reduction from the problem of inferring a small set of explaining variables to the problem of finding a small cover of E by sets from $E_1, \ldots, E_n$. This allows us to use algorithms for the set cover problem to find explaining variables. The best known and probably most generic algorithm for this problem is a greedy algorithm that successively picks a set that covers the largest amount of elements not covered so far. This algorithm, which we call Greedy, is presented as Algorithm 4.1.

Each iteration of the while-loop in lines 4 to 9 of Greedy is referred to as a *round* of Greedy. If there are several sets of maximum cardinality in line 7 of Greedy, it may pick one of them at random (or follow any fixed selection rule such as picking the set with the smallest index). It is natural to call Greedy (or any algorithm that learns relevant attributes) *successful* on input S if Greedy$(S) = \mathrm{rel}(f)$. In some situations, however, one may be content with finding a superset of the relevant variables that is at most a constant factor larger than $\mathrm{rel}(f)$. A corresponding notion of success for Greedy is captured as follows.

Definition 4.1.4 ($\boldsymbol{\lambda}$-success). Let $f : \{0,1\}^n \to \{0,1\}$, S be a sample for f, and $\lambda \geq 1$. Greedy is λ-*successful* on input S if and only if

1. $|$Greedy$(S)| \leq \lambda \cdot |\,\mathrm{rel}(f)|$ and

2. Greedy$(S) \supseteq \mathrm{rel}(f)$.

Greedy is *successful* (or *succeeds*) if and only if it is 1-*successful*, i.e.,

$$\textsc{Greedy}(S) = \mathrm{rel}(f) \ ,$$

otherwise we say that it *fails*. Greedy λ-*fails* if and only if it is not λ-successful.

Example 4.1.5. Let $f : \{0,1\}^6 \to \{0,1\}$ be defined by

$$f(x_1, \dots, x_6) = x_2 \oplus x_4 \ .$$

Then the sample S given in Table 4.1 is a sample for f. If we assume that during the execution of Greedy, in case of equal set sizes, the algorithm picks the set with the smallest index, then Greedy outputs x_1, x_2, and x_3 (in this order). Hence, it fails on input S. Even worse, it λ-fails for all $\lambda \geq 1$ since it does not output x_4.

4.2 Key Lemmas for the Algorithm Analyses

In this section, we provide three key lemmas that will be used in the proofs of our main results in Section 4.3. Recall from Definition 2.2.2 that the expanded variable space contains the variables

$$x_I = \bigoplus_{i \in I} x_i \quad \text{for } I \subseteq [n] \ .$$

Define the corresponding edge sets

$$E_I = \left\{ \{k, \ell\} \in E \mid k, \ell \in [m], x_I^k \neq x_I^\ell \right\} \ .$$

Since x_I^k and x_I^ℓ differ if and only if the number of $i \in I$ with $x_i^k \neq x_i^\ell$ is odd, we obtain that $E_I = \triangle_{i \in I} E_i$, where $\triangle$ denotes the symmetric difference.

Suppose that Greedy has put the variables $x_{i_1}, \dots, x_{i_s}$ into R after s rounds. Hence, all edges in $E' = E_{i_1} \cup \cdots \cup E_{i_s}$ have been covered. The number of remaining edges that can be covered by variable x_i in the next round is $|E_i \setminus E'|$. Provided that $x_{i_1}, \dots, x_{i_s}$ are all relevant, we would like to estimate the set size $|E_i \setminus E'|$ in dependence of properties of f. As we do not see any direct way of doing so, we take a detour via the cardinalities of the sets E_I. These turn out to be approximable quite efficiently, as we will show in Lemma 4.2.2. But let us first show how to express the cardinality of $E_i \setminus E'$ in terms of the cardinalities of the sets E_I, $I \subseteq \{i_1, \dots, i_s\}$:

Lemma 4.2.1. *Let $S \in (\{0,1\}^n \times \{0,1\})^m$ be a sample and $G_S = (V, E)$ be the corresponding functional relations graph. Let $R \subsetneq [n]$ and $i^* \in [n] \setminus R$ and define $E' = \bigcup_{i \in R} E_i$. Then*

$$|E_{i^*} \setminus E'| = 2^{-|R|} \sum_{I \subseteq R} (|E_{I \cup \{i^*\}}| - |E_I|) \ .$$

Proof. Recall the definition of the characteristic vector $c(e) \in \{0,1\}^n$ of an edge $e \in E$ given in Equation (4.1). Using the notation of the expanded variable space, we write

$$c_I(e) = \bigoplus_{i \in I} c_i(e) \ .$$

Clearly, we can write $|E_I| = \sum_{e \in E} c_I(e)$ for all $I \subseteq [n]$. Let $I \subseteq R$. Since for $e \in E \setminus E_{i^*}$, we have $c_{I \cup \{i^*\}}(e) = c_I(e)$,

$$|E_{I \cup \{i^*\}}| - |E_I| = \sum_{e \in E_{i^*}} \left(c_{I \cup \{i^*\}}(e) - c_I(e)\right) \ .$$

For $e \in E_{i^*}$, we have

$$c_{I \cup \{i^*\}}(e) - c_I(e) = \begin{cases} -1 & \text{if } e \in E_I \ , \\ 1 & \text{if } e \notin E_I \ . \end{cases}$$

For all edges $e \in E'$, we have $e \in E_I$ for exactly half of the sets $I \subseteq R$. To see this, let $e \in E'$, i.e., $e \in E_j$ for some $j \in R$. Then for each set $I \subseteq R \setminus \{j\}$, we have $e \in E_I$ if and only if $e \notin E_{I \cup \{j\}}$. Therefore, for such edges,

$$\sum_{I \subseteq R} (c_{I \cup \{i^*\}}(e) - c_I(e)) = 0 \ .$$

Consequently,

$$\begin{aligned} \sum_{I \subseteq R} (|E_{I \cup \{i^*\}}| - |E_I|) &= \sum_{e \in E_{i^*}} \sum_{I \subseteq R} (c_{I \cup \{i^*\}}(e) - c_I(e)) \\ &= \sum_{e \in E_{i^*} \setminus E'} \sum_{I \subseteq R} (c_{I \cup \{i^*\}}(e) - c_I(e)) \ . \end{aligned}$$

On the other hand, if $e \in E_{i^*} \setminus E'$, then $e \notin E_I$ for all $I \subseteq R$. In this case,

$$\sum_{I \subseteq R} (c_{I \cup \{i^*\}}(e) - c_I(e)) = 2^{|R|} \ ,$$

and thus

$$\sum_{I \subseteq R} (|E_{I \cup \{i^*\}}| - |E_I|) = 2^{|R|} \cdot |E_{i^*} \setminus E'| \ ,$$

proving the claim. □

Now we are concerned with the estimation of the cardinalities $|E_I|$, $I \subseteq [n]$. For $a, b \in \{0,1\}$, let

$$\alpha_I^{ab} = \Pr[x_I = a \wedge f(x) = b] \ , \tag{4.2}$$

where $x \in \{0,1\}^n$ is drawn according to the uniform distribution. It follows that $\alpha_I^{a0} + \alpha_I^{a1} = \Pr[x_I = a] = 1/2$ for $I \neq \emptyset$ and $\alpha_I^{0b} + \alpha_I^{1b} = \Pr[f(x) = b]$ for all $I \subseteq [n]$.

A (noise-free) sample of size m consists of the outcomes of m independent draws of $x^k \in \{0,1\}^n$ and the corresponding classifications $y^k = f(x^k) \in \{0,1\}$. In the following, all probabilities and expectations are taken with respect to the random experiment of "drawing a sample of size m" for an arbitrary but fixed m. For all $I \subseteq [n]$ and all pairs of example indices $k, \ell \in [m]$ with $k \neq \ell$, the probability that $\{k,\ell\} \in E_I$ is

$$\Pr[x_I^k \neq x_I^\ell \wedge y^k \neq y^\ell] = 2(\alpha_I^{00}\alpha_I^{11} + \alpha_I^{10}\alpha_I^{01}) \ .$$

Since there are $\frac{1}{2}(m-1)m$ such pairs, the expectation of $|E_I|$ is $\alpha_I(m-1)m$ with

$$\alpha_I = \alpha_I^{00}\alpha_I^{11} + \alpha_I^{10}\alpha_I^{01} \ . \tag{4.3}$$

Next we prove a Chernoff style mass concentration for the cardinalities $|E_I|$. It shows that for a sufficiently large sample size, $|E_I|$ is likely to be very close to $\alpha_I \cdot m^2$.

Lemma 4.2.2. *There exist $c_1, c_2 > 0$ such that for every $f : \{0,1\}^n \to \{0,1\}$, given a uniformly distributed sample S of size m for f, for all $I \subseteq [n]$ and arbitrary $\epsilon \in [0,1]$,*

$$\Pr\left[\left||E_I| - \alpha_I m^2\right| > \epsilon m^2\right] \ < \ c_1 e^{-c_2\epsilon^2 m} \ .$$

Proof. Let S be a uniformly distributed sample of size m. For $a, b \in \{0,1\}$, let A_I^{ab} denote the set of example indices k such that $(x_I^k, y^k) = (a,b)$. In E_I, there are edges between all pairs (x^k, y^k) and (x^ℓ, y^ℓ) with $x_I^k \neq x_I^\ell$ and $y^k \neq y^\ell$. Thus, we obtain

$$E_I = \{\{k,\ell\} \mid k \in A_I^{a,0}, \ell \in A_I^{1-a,1}, a \in \{0,1\}\}$$

and hence,

$$|E_I| = |A_I^{00}| \cdot |A_I^{11}| + |A_I^{10}| \cdot |A_I^{01}| \quad .$$

The expected number of examples with $x_I^k = a$ and $y^k = b$ clearly is $\alpha_I^{ab} m$. By the Hoeffding bound (Lemma 3.1.2), with probability at least $1 - 2e^{-2\delta^2 m}$,

$$|A_I^{ab} - \alpha_I^{ab} m| \leq \delta m \ . \tag{4.4}$$

Moreover, (4.4) holds for all $a, b \in \{0,1\}$ simultaneously with probability at least

$1-8e^{-2\delta^2 m}$. In this case,

$$\begin{array}{rcl}\big||E_I|-\alpha_I m^2\big| & = & \big||A_I^{00}|\cdot|A_I^{11}|+|A_I^{10}|\cdot|A_I^{01}|-(\alpha_I^{00}\alpha_I^{11}+\alpha_I^{10}\alpha_I^{01})m^2\big| \\ & \leq & \big||A_I^{00}|\cdot|A_I^{11}|-\alpha_I^{00}m\cdot|A_I^{11}|\big|+\big|\alpha_I^{00}m\cdot|A_I^{11}|-\alpha_I^{00}m\cdot\alpha_I^{11}m\big| \\ & & +\big||A_I^{10}|\cdot|A_I^{01}|-\alpha_I^{10}m\cdot|A_I^{01}|\big|+\big|\alpha_I^{10}m\cdot|A_I^{01}|-\alpha_I^{10}m\cdot\alpha_I^{01}m\big| \\ & \leq & \big||A_I^{00}|-\alpha_I^{00}m\big|\cdot|A_I^{11}|+\big||A_I^{11}|-\alpha_I^{11}m\big|\cdot\alpha_I^{00}m \\ & & +\big||A_I^{10}|-\alpha_I^{10}m\big|\cdot|A_I^{01}|+\big||A_I^{01}|-\alpha_I^{01}m\big|\cdot\alpha_I^{10}m \\ & \leq & \delta m\big(|A_I^{11}|+\alpha_I^{00}m+|A_I^{01}|+\alpha_I^{10}m\big) \\ & \leq & \delta m\big((\alpha_I^{11}+\delta)m+\alpha_I^{00}m+(\alpha_I^{01}+\delta)m+\alpha_I^{10}m\big) \\ & \leq & \delta(1+2\delta)m^2\ .\end{array}$$

Thus, we can find $c_1, c_2 > 0$ such that

$$\big||E_I|-\alpha_I m^2\big| \ < \ c_1 e^{-c_2\epsilon^2 m}$$

as claimed ($c_1 = 8$, $c_2 = 1/2$, and $\delta = \epsilon/2$ do the job). □

Before stating the third lemma, let us briefly take a closer look at the cardinalities $|E_i|$ for irrelevant variables x_i. Since for these, the value of x_i is independent of the classification $f(x)$, $\alpha_i^{ab} = \frac{1}{2}\Pr[f(x) = b]$. Consequently, $\alpha_i = \frac{1}{2}\Pr[f(x) = 0]\Pr[f(x) = 1] = \frac{1}{2}\operatorname{Var}[f]$. Hence, the expectation of $|E_i|$ is $\frac{1}{2}\operatorname{Var}[f]m(m-1) \approx \frac{1}{2}\operatorname{Var}[f]m^2$. A moment's reflection shows that also for $I \subseteq [n]$ with $I \not\subseteq \operatorname{rel}(f)$, the expectation of $|E_I|$ is equal to $\frac{1}{2}\operatorname{Var}[f]m(m-1)$.

The following lemma generalizes this result to *arbitrary* $I \subseteq [n]$, revealing an unexpected relationship between the cardinalities $|E_I|$ and the Fourier coefficients $\hat{f}(I)$. Recall that for $I \subseteq [n]$ with $I \not\subseteq \operatorname{rel}(f)$, $\hat{f}(I) = 0$ by Lemma 2.3.4.

Lemma 4.2.3. *Let $f : \{0,1\}^n \to \{0,1\}$ and $I \subseteq [n]$ with $I \neq \emptyset$. Then*

$$\alpha_I = \tfrac{1}{2}\Big(\operatorname{Var}[f] + \hat{f}(I)^2\Big)\ . \tag{4.5}$$

Proof. From

$$\hat{f}(I) = 2^{-n}\sum_{x\in\{0,1\}^n} f(x)\cdot(-1)^{x_I} = \alpha_I^{01} - \alpha_I^{11}\ , \tag{4.6}$$

it follows that

$$\begin{array}{rcl}\alpha_I - \frac{1}{2}\operatorname{Var}[f] & = & \alpha_I^{00}\alpha_I^{11}+\alpha_I^{10}\alpha_I^{01}-\frac{1}{2}\operatorname{Var}[f] \\ & = & (\frac{1}{2}-\alpha_I^{01})\alpha_I^{11}+(\frac{1}{2}-\alpha_I^{11})\alpha_I^{01}-\frac{1}{2}\Pr[f(x)=0]\Pr[f(x)=1] \\ & = & \frac{1}{2}\,(\alpha_I^{11}+\alpha_I^{01}-4\alpha_I^{01}\alpha_I^{11}-\Pr[f(x)=1]+\Pr[f(x)=1]^2) \\ & = & \frac{1}{2}\Pr[f(x)=1]^2-2\alpha_I^{01}\alpha_I^{11} \\ & = & \frac{1}{2}(\alpha_I^{01}-\alpha_I^{11})^2 \ = \ \frac{1}{2}\hat{f}(I)^2\ .\end{array}$$

□

Note that always $x_\emptyset = 0$, independent of the values of $x_1, \dots, x_n$. Thus, $E_\emptyset = \emptyset$ and $\alpha_\emptyset = 0$. In particular, (4.5) is not valid for $I = \emptyset$.

4.3 Analysis of the Greedy Algorithm

4.3.1 Greedy Succeeds for all Functions that are Fourier-accessible

In this section, we state and prove our main results for the Greedy algorithm. Let us start with the positive result.

Theorem 4.3.1. *There is a polynomial p such that the following holds. Let $f : \{0,1\}^n \to \{0,1\}$ be a Fourier-accessible concept, $d = |\operatorname{rel}(f)|$, and $\delta > 0$. Let S be a uniformly distributed sample for f of size*

$$m \geq p(2^d, \log n, \log(1/\delta)) \ .$$

Then $\textsc{Greedy}(S) = \operatorname{rel}(f)$ *with probability at least* $1 - \delta$.

Proof. We first show that with high probability, Greedy outputs *at least d* variables, provided that m is sufficiently large. By Lemma 3.5.1, with probability at least $1 - \delta/2$, *any* d-junta that is consistent with a sample S for f of size $m \geq m_0 = 2^{2d+1} \ln \frac{2n}{\delta}$ must be f itself. Thus, with probability at least $1 - \delta/2$, E cannot be covered by less than d sets E_i since such a covering would yield a consistent concept that depends on strictly less than d variables.

Now assume that Greedy indeed outputs at least d variables. Let the sequence of variables output by Greedy start with $x_{i_1}, \dots, x_{i_d}$. For $s \in [d]$, let $R_s = \{i_1, \dots, i_s\}$. We prove that with high probability, each variable that is output is relevant to f. This implies that Greedy halts after exactly d rounds since E can always be covered by the sets E_i, $x_i \in \operatorname{rel}(f)$.

Let $\epsilon = 2^{-3d-3}$. For each $I \subseteq [n]$ with $1 \leq |I| \leq d$, we have

$$\Pr\left[\left||E_I| - \alpha_I m^2\right| > \epsilon m^2\right] < c_1 e^{-c_2 \epsilon^2 m}$$

for some constants $c_1, c_2 > 0$ by Lemma 4.2.2. Consequently,

$$\forall I \subseteq [n] \text{ such that } 1 \leq |I| \leq d : \ \left||E_I| - \alpha_I m^2\right| \leq \epsilon m^2 \tag{4.7}$$

with probability at least $\rho = 1 - n^d \cdot c_1 e^{-c_2 \epsilon^2 m}$ (since $V(n,d) - 1 \leq n^d$). In the following, we assume that (4.7) holds. Thus, all subsequent consequences of (4.7) hold with probability at least ρ.

We show by induction on $s \in [d]$ that $R_s \subseteq \mathrm{rel}(f)$. For $s = 0$,

$$R_0 = \emptyset \subseteq \mathrm{rel}(f) \ .$$

For the induction step, let $s \in \{0, \ldots, d-1\}$ and assume that $R_s \subseteq \mathrm{rel}(f)$. For $i \in [n] \setminus R_s$, denote by $E_i^{(s)}$ the set of remaining edges in E_i after the s^{th} round of GREEDY, i.e., $E_i^{(s)} = E_i \setminus \{E_{i_1} \cup \cdots \cup E_{i_s}\}$.

Our goal is to show that there exists a relevant variable x_{i^*} such that $E_{i^*}^{(s)}$ is larger than $E_j^{(s)}$ for all irrelevant variables x_j. Since we have not found all relevant variables after round s, there is an $i^* \in \mathrm{rel}(f) \setminus R_s$ and an $I_0 \subseteq R_s$ such that $\hat{f}(I_0 \cup \{i^*\}) \neq 0$ and hence $|\hat{f}(I_0 \cup \{i^*\})| \geq 2^{-d}$ by Lemma 2.3.6. If there were no such i^* and I_0, then none of the variables x_i with $i \in \mathrm{rel}(f) \setminus R_s$ would be accessible, contradicting the assumption that f is Fourier-accessible. For arbitrary $x_j \in \mathrm{irrel}(f)$, Lemma 4.2.1 implies

$$|E_{i^*}^{(s)}| - |E_j^{(s)}| = 2^{-s} \sum_{I \subseteq R_s} (|E_{I \cup \{i^*\}}| - |E_{I \cup \{j\}}|) \ .$$

From (4.7), we obtain

$$|E_{i^*}^{(s)}| - |E_j^{(s)}| \geq 2^{-s} \sum_{I \subseteq R_s} \left((\alpha_{I \cup \{i^*\}} - \epsilon) m^2 - (\alpha_{I \cup \{j\}} + \epsilon) m^2 \right) \ .$$

Now, by Lemma 4.2.3,

$$\begin{aligned} |E_{i^*}^{(s)}| - |E_j^{(s)}| \ \geq \ & 2^{-s} \sum_{I \subseteq R_s} \left[\left(\tfrac{1}{2} \mathrm{Var}[f] + \tfrac{1}{2} \hat{f}(I \cup \{i^*\})^2 - \epsilon \right) m^2 \right. \\ & \qquad \left. - \left(\tfrac{1}{2} \mathrm{Var}[f] + \tfrac{1}{2} \hat{f}(I \cup \{j\})^2 + \epsilon \right) m^2 \right] \\ \geq \ & 2^{-s} \sum_{I \subseteq R_s} \left(\tfrac{1}{2} \hat{f}(I \cup \{i^*\})^2 - \tfrac{1}{2} \hat{f}(I \cup \{j\})^2 - 2\epsilon \right) m^2 \ . \end{aligned}$$

Since $x_j \in \mathrm{irrel}(f)$, it follows that $\hat{f}(I \cup \{j\}) = 0$ by Lemma 2.3.5. Thus,

$$\begin{aligned} |E_{i^*}^{(s)}| - |E_j^{(s)}| \ \geq \ & 2^{-s} \sum_{I \subseteq R_s} \left(\tfrac{1}{2} \hat{f}(I \cup \{i^*\})^2 - 2\epsilon \right) m^2 \\ \geq \ & 2^{-s} \left(\tfrac{1}{2} \hat{f}(I_0 \cup \{i^*\})^2 + 2^s \cdot (-2\epsilon) \right) m^2 \\ \geq \ & (2^{-3d-1} - 2\epsilon) m^2 \ \geq \ 2^{-3d-2} m^2 > 0 \ . \end{aligned}$$

Consequently, in round $s+1$, GREEDY prefers the relevant variable x_{i^*} to all irrelevant variables. In particular, GREEDY selects some relevant variable in round $s+1$.

It suffices to choose $m \geq m_1 = c_2^{-1} \cdot 2^{6d-6}(d \ln n + \ln(2c_1/\delta))$ to have $\rho \geq 1 - \delta/2$. In total, we can choose $m = \max\{m_0, m_1\}$, which is polynomial in 2^d, $\log n$, and $\log(1/\delta)$, to guarantee that Greedy outputs exactly the relevant variables of f. □

Example 4.3.2. The function $f_1 : \{0,1\}^3 \to \{0,1\}$ in Table 2.1 is Fourier-accessible (see Example 2.4.8). By Theorem 4.3.1, for any function $f : \{0,1\}^n \to \{0,1\}$ that has f_1 as its base function (see Definition 2.3.2), Greedy succeeds with probability at least $1-\delta$ for sample size polynomial in 2^d, $\log n$, and $\log(1/\delta)$.

4.3.2 Greedy Fails for all Functions that are not Fourier-accessible

If a concept is not Fourier-accessible, then one of its relevant variables is not accessible. The proof of Theorem 4.3.1 shows that Greedy first outputs all accessible variables with high probability. Once all of these have been output, the intuition is that the non-accessibility of the other relevant variables makes them statistically indistinguishable from the irrelevant variables. In particular, each inaccessible but relevant variable will be selected by Greedy with the same probability as each irrelevant variable. Assuming that the number of irrelevant variables is much larger than the number of relevant ones, it becomes very likely that Greedy picks an irrelevant variable and thus fails.

Before we prove our second main result for the Greedy algorithm, we start with a lemma that makes the above idea precise.

Lemma 4.3.3. *Let $f : \{0,1\}^n \to \{0,1\}$ be a concept that is not Fourier-accessible and S be a sample for f of arbitrary size m. Let $x_{i_1}, \ldots, x_{i_t}$ be the variables output by* Greedy *on input S. Let $s \in \{0, \ldots, t-1\}$ and define $E_i^{(s)} = E_i \setminus (E_{i_1} \cup \cdots \cup E_{i_s})$ for $i \in [n]$. Given that*

$$\{x_{i_1}, \ldots, x_{i_s}\} \subseteq \operatorname{acc}(f) \cup \operatorname{irrel}(f) ,$$

the following statements hold.

(a) Let x_i be a variable that is relevant but not accessible. Then the random variables $|E_i^{(s)}|$ and all $|E_j^{(s)}|$, $x_j \in \operatorname{irrel}(f)$, conditional to any fixed values of $x_{i_1}^k, \ldots, x_{i_s}^k$ and $y^k = f(x^k)$, $k \in [m]$, are independent and identically distributed.

(b) The probability that $x_{i_{s+1}}$ is relevant to f but not accessible is at most

$$\frac{|\operatorname{rel}(f) \ \cap \ \operatorname{inacc}(f)|}{|\operatorname{irrel}(f) \setminus \{x_{i_1}, \ldots, x_{i_s}\}|} .$$

Proof. (a) Let $R = \{i_1, \ldots, i_s\}$. We show that for $a, b \in \{0,1\}$ and $v \in \{0,1\}^R$ such that there exists $x \in \{0,1\}^n$ with $x|_R = v$ and $f(x) = b$,

$$\Pr[x_i = a \mid x|_R = v \wedge f(x) = b] = \tfrac{1}{2}\ ,$$

just as for the irrelevant variables: Let $I \subseteq R$. If $I \subseteq \mathrm{acc}(f)$, then since x_i is not accessible, $\hat{f}(I \cup \{i\}) = 0$. If $I \not\subseteq \mathrm{acc}(f)$, then I contains some irrelevant variable index, and hence $\hat{f}(I \cup \{i\}) = 0$ by Lemma 2.3.4. By Corollary 2.2.5, $\widehat{f_v}(i) = 0$ for all $v \in \{0,1\}^R$. From

$$\begin{aligned} \widehat{f_v}(i) &= \Pr[x_i = 0 \wedge f_v(x) = 1] - \Pr[x_i = 1 \wedge f_v(x) = 1] \\ &= \Pr[x_i = 1 \wedge f_v(x) = 0] - \Pr[x_i = 0 \wedge f_v(x) = 0]\ , \end{aligned} \tag{4.8}$$

and

$$\Pr[x_i = 0 \wedge f_v(x) = b] + \Pr[x_i = 1 \wedge f_v(x) = b] = \Pr[f_v(x) = b]\ ,$$

we can deduce $\Pr[x_i = a \wedge f_v(x) = b] = \frac{1}{2}\Pr[f_v(x) = b]$. Consequently, we obtain

$$\begin{aligned} \Pr[x_i = a \wedge x|_R = v \wedge f(x) = b] &= \Pr\big[x_i = a \wedge f(x) = b \bigm| x|_R = v\big] \cdot \Pr[x|_R = v] \\ &= \Pr[x_i = a \wedge f_v(x) = b] \cdot \Pr[x|_R = v] \\ &= \tfrac{1}{2}\Pr[f_v(x) = b] \cdot \Pr[x|_R = v] \\ &= \tfrac{1}{2}\Pr\big[f(x) = b \bigm| x|_R = v\big] \cdot \Pr[x|_R = v] \\ &= \tfrac{1}{2}\Pr[f(x) = b \wedge x|_R = v]\ . \end{aligned}$$

Thus, $\Pr\big[x_i = a \bigm| x|_R = v \wedge f(x) = b\big] = 1/2$, which proves the claim.

As a consequence of the latter, conditional to the values of $x^k_{i_1}, \ldots, x^k_{i_s}$ and $f(x^k)$, $k \in [m]$, the cardinalities $|E^{(s)}_i|$ and all $|E^{(s)}_j|$, $x_j \notin \mathrm{rel}(f)$, are identically distributed (since these cardinalities only depend on the outcomes of x^k_i, $f(x^k)$, and $x^k_{i_1}, \ldots, x^k_{i_s}$, $k \in [m]$, and since all examples are drawn independently). The independence is obvious.

(b) For a fixed variable x_i that is relevant but not accessible, the probability that $x_{i_{s+1}} = x_i$ is at most as large as the probability that $x_{i_{s+1}} = x_i$ conditional to $x_{i_{s+1}} \in \{x_i\} \cup (\mathrm{irrel}(f) \setminus \{x_{i_1}, \ldots, x_{i_s}\})$. Since all cardinalities $E^{(s)}_j$ corresponding to the variables in $\{x_i\} \cup (\mathrm{irrel}(f) \setminus \{x_{i_1}, \ldots, x_{i_s}\})$ are identically distributed, the probability that x_i is selected in round $s+1$ is at most

$$\frac{1}{|\,\mathrm{irrel}(f) \setminus \{x_{i_1}, \ldots, x_{i_s}\}| + 1}\ .$$

Hence, the probability that $x_{i_{s+1}}$ is relevant but not accessible is at most

$$\frac{|\,\mathrm{rel}(f)\ \cap\ \mathrm{inacc}(f)|}{|\,\mathrm{irrel}(f) \setminus \{x_{i_1}, \ldots, x_{i_s}\}|}\ .$$

□

The following negative result complements Theorem 4.3.1.

Theorem 4.3.4. *Let $f : \{0,1\}^n \to \{0,1\}$ be a concept that is not Fourier-accessible, $d = |\operatorname{rel}(f)|$, and $\lambda \geq 1$. Given a uniformly distributed sample S for f of arbitrary size,* Greedy *λ-fails on input S with probability at least*

$$1 - \frac{\lambda d^2}{n - \lambda d} .$$

Proof. Let $x_{i_1}, \ldots, x_{i_t}$ be the variables output by Greedy. For fixed $s \in \{0, \ldots, t-1\}$, the probability that $\{x_{i_1}, \ldots, x_{i_s}\} \subseteq \operatorname{acc}(f) \cup \operatorname{irrel}(f)$ *and* that $x_{i_{s+1}}$ is relevant but not accessible is at most as large as the probability that $x_{i_{s+1}}$ is relevant but not accessible *conditional to* $\{x_{i_1}, \ldots, x_{i_s}\} \subseteq \operatorname{acc}(f) \cup \operatorname{irrel}(f)$. This is at most $d/(|\operatorname{irrel}(f) \setminus \{x_{i_1}, \ldots, x_{i_s}\}|)$ by Lemma 4.3.3 (b).

Suppose that Greedy is λ-successful on input S, i.e.,

$$t \leq \lambda \cdot d \text{ and } \{x_{i_1}, \ldots, x_{i_t}\} \supseteq \operatorname{rel}(f) .$$

Since f is not Fourier-accessible, there exists $s \in \{0, \ldots, t-1\}$ such that $x_{i_1}, \ldots, x_{i_s} \in \operatorname{acc}(f) \cup \operatorname{irrel}(f)$ and $x_{i_{s+1}}$ is relevant but not accessible. The probability for the latter event is at most $t \cdot \frac{d}{n-t}$. Hence, the probability that Greedy fails is at least $1 - \frac{\lambda d^2}{n - \lambda d}$. □

Corollary 4.3.5. *Let $p_\lambda(n,d)$ denote the probability that for any given concept $f : \{0,1\}^n \to \{0,1\}$ with $|\operatorname{rel}(f)| = d$ that is* not *Fourier-accessible and for any uniformly distributed sample S for f,* Greedy *λ-fails. Then for fixed $\lambda \geq 1$,*

(a) for fixed d, $\lim_{n\to\infty} p_\lambda(n,d) = 1$ and

(b) for $d \to \infty$ and $n = n(d) \in \omega(d^2)$, $\lim_{d\to\infty} p_\lambda(n,d) = 1$.

Example 4.3.6. The function $f_2 : \{0,1\}^3 \to \{0,1\}$ in Table 2.1 is not Fourier-accessible. By Theorem 4.3.4, for any function $f : \{0,1\}^n \to \{0,1\}$ that has f_2 as its base function (see Definition 2.3.2), Greedy fails with probability at least $1 - \frac{9}{n-3}$.

Note that Theorem 4.3.4 not only says that Greedy fails (with high probability) for concepts that are not Fourier-accessible, but that Greedy even fails to find all relevant variables of the target concept in $\lambda \cdot |\operatorname{rel}(f)|$ rounds for any $\lambda \geq 1$. In addition, note that the claim in Theorem 4.3.4 is independent of the sample size.

In the literature, it has often been emphasized that Greedy has a “logarithmic approximation guarantee” (see [AB96, AMK03, BL97, FA05]), i.e., given a sample S for f of size m, Greedy finds a set of at most $(2 \ln m + 1) \cdot |\operatorname{rel}(f)|$

variables that explain S. Theorem 4.3.4 shows that if f is not Fourier-accessible, then with probability at least

$$1 - \frac{(2\ln m + 1)d^2}{n - (2\ln m + 1)d} \ ,$$

these variables *do not contain all relevant variables* (where $d = |\operatorname{rel}(f)|$). Thus, given a sample of a target concept that is not Fourier accessible, GREEDY *misses* some relevant variable with high probability, provided that $m \in 2^{o(n)}$. In other words, the positive approximability properties of the greedy strategy for the SET COVER problem do not translate to the learning situation. The fact that GREEDY outputs at most $(2\ln m \ + 1) \cdot |\operatorname{rel}(f)|$ variables only guarantees that any sample of size m can be explained by this amount of variables.

4.3.3 Non-uniform Attribute Distributions

Let $D : \{0,1\}^n \to [0,1]$ be a non-uniform attribute distribution. Unfortunately, while Lemma 4.2.3 is also valid with α_I defined with respect to D, it does not seem to be possible to show an equation in the spirit of (4.5) and present a characterization of the concept class for which GREEDY works under this assumption. In fact, it is easy to find functions f and product distributions D with rates $d_1, \dots, d_n$ such that

(a) there are $x_i, x_j \in \operatorname{irrel}(f)$ such that the expected sizes of E_i and E_j differ or

(b) there are $x_i \in \operatorname{rel}(f)$ and $x_j \in \operatorname{irrel}(f)$ such that the expected sizes of E_i and E_j are equal.

For item (a), we simply pick different d_i and d_j. Then $\alpha_i \neq \alpha_j$ whenever $\Pr[f(x) = 0] \neq 0 \neq \Pr[f(x) = 1]$.

For item (b), consider $f \colon \{0,1\}^3 \to \{0,1\}$ defined by $f(x_1, x_2, x_3) = x_1 \wedge x_2$. Let $d_1 = 2/3$, $d_2 = 1/2$, and $d_3 = 1/2$. Then we have $\alpha_1 = \alpha_3 = 1/9$. Variable x_1 is relevant to f, but x_3 is not.

Another example that shows that different methods are needed for different distributions will be presented at the end of Section 4.7.

4.4 Extension of the Greedy Algorithm to Larger Concept Classes

As we have seen in Example 2.4.8 (d), the *not-all-equal* function

$$\text{NAE} : \{0,1\}^d \to \{0,1\}$$

Algorithm 4.2 τ-Greedy.

```
input $S = ((x_1^k, \ldots, x_n^k), y^k)_{k \in [m]}$
$E \leftarrow \{\{k, \ell\} \mid k, \ell \in [m], y^k \neq y^\ell\}$
$R \leftarrow \emptyset$
while $E \neq \emptyset$ do
  for $I \subseteq [n]$ with $1 \leq |I| \leq \tau$ do
    $E_I \leftarrow \{\{k, \ell\} \in E \mid x_I^k \neq x_I^\ell\}$
  select $I \subseteq [n]$, $1 \leq |I| \leq \tau$, with maximum $|E_I|$
  $E \leftarrow E \setminus \bigcup_{i \in I} E_i$
  $R \leftarrow R \cup \{x_i \mid i \in I\}$
output $\tau$-Greedy$(S) = R$
```

is 2-low but not 1-low. By Lemma 2.3.4, for concepts which restricted to their relevant variables become equal to NAE, it suffices to check for all $I \subseteq [n]$ with $|I| = 2$, whether $\hat{f}(I) \neq 0$. This motivates us to seek for an extension of the greedy approach that is also able to cope with τ-low juntas for $\tau > 1$.

In this section, we show that allowing Greedy to choose from the sets E_I, $1 \leq |I| \leq \tau$, the algorithm can cope exactly with the τ-Fourier-accessible concepts, where $\tau \in [d]$ is some fixed parameter. The corresponding algorithm, which we call τ-Greedy, is presented as Algorithm 4.2. Note that 1-Greedy matches Greedy. The running time of τ-Greedy is $\binom{n}{\tau}$ times a polynomial in m and n, which is dominated by $n^\tau \cdot \mathrm{poly}(m, n)$.

As one would expect, the basic ideas underlying the proofs presented in this section are similar to those in Section 4.3. However, we have to carefully find the correct formulations of the results and proofs, which are not immediate in all places, in particular concerning the negative result for τ-Greedy (see Lemma 4.4.3 and Theorem 4.4.4).

We now extend the results from Section 4.3 to τ-Greedy and τ-Fourier accessible concepts. In particular, we need to generalize Lemma 4.2.1 to use it in the proofs of the extended results.

Lemma 4.4.1. *Let $S \in (\{0,1\}^n \times \{0,1\})^m$ be a sample and $G_S = (V, E)$ be the corresponding functional relations graph. Let $R \subsetneq [n]$ and $I^* \subseteq [n] \setminus R$ and define $E' = \bigcup_{i \in R} E_i$. Then*

$$|E_{I^*} \setminus E'| = 2^{-|R|} \sum_{I \subseteq R} (|E_{I \cup I^*}| - |E_I|) \ .$$

Proof. Exactly as the proof of Lemma 4.2.1, except that every occurrence of $\{i^*\}$ has to be substituted by I^*. □

Theorem 4.4.2. *There is a polynomial p such that the following holds. Let $f : \{0,1\}^n \to \{0,1\}$ be a τ-Fourier-accessible concept, $1 \le \tau \le d = |\operatorname{rel}(f)|$, and let $\delta > 0$. Let S be a uniformly distributed sample S for f of size*

$$m \ge p(2^d, \log n, \log(1/\delta)) .$$

*Then τ-*GREEDY*$(S) = \operatorname{rel}(f)$ with probability at least $1 - \delta$.*

Proof. The proof is very similar to that of Theorem 4.3.1, so we only point out the differences. Instead of having an $i^* \in \operatorname{rel}(f) \setminus R_s$ and an $I_0 \subseteq R_s$ such that $\hat{f}(I_0 \cup \{i^*\}) \neq 0$, we can only guarantee that there are $i^* \in \operatorname{rel}(f) \setminus R_s$, $I^* \subseteq [n] \setminus R_s$ with $i^* \in I^*$ and $|I^*| \le \tau$, and $I_0 \subseteq R_s$ such that $\hat{f}(I_0 \cup I^*) \neq 0$ and hence $|\hat{f}(I_0 \cup I^*)| \ge 2^{-d}$. For arbitrary $J \subseteq [n] \setminus R_s$ with $J \cap \operatorname{irrel}(f) \neq \emptyset$ and $|J| \le \tau$, Lemma 4.4.1 implies

$$|E_{I^*}^{(s)}| - |E_J^{(s)}| = 2^{-s} \sum_{\emptyset \subseteq I \subseteq R_s} (|E_{I \cup I^*}| - |E_{I \cup J}|) > 0 ,$$

where the detailed calculation is the same as in the proof of Theorem 4.3.1, except that all occurrences of $\{i^*\}$ have to be substituted by I^* and all occurrences of $\{j\}$ have to be substituted by J.

Consequently, in round $s+1$, GREEDY prefers the set E_{I^*} (with $I^* \subseteq \operatorname{rel}(f)$) to all sets E_J with $J \cap \operatorname{irrel}(f) \neq \emptyset$. In particular, GREEDY adds only relevant variables to its output in round $s + 1$. □

To prove that τ-GREEDY fails for all concepts that are not τ-Fourier-accessible, we need to cast (and prove) an analog of Lemma 4.3.3 for such concepts.

Lemma 4.4.3. *Let $f : \{0,1\}^n \to \{0,1\}$ be a concept that is not τ-Fourier-accessible for some $\tau \in [n]$ and S be a uniformly distributed sample for f of arbitrary size m. Let t denote the number of rounds τ-*GREEDY *runs on input S and denote by I_s, $s \in [t]$, the set I selected in line 7 of τ-*GREEDY *in round s. Fix $s \in \{0, \ldots, t-1\}$ and define*

$$E_I^{(s)} = E_I \setminus \bigcup_{r=1}^{s} \bigcup_{i \in I_r} E_i$$

*for $I \subseteq [n]$. Let $x_{i_1}, \ldots, x_{i_r}$ be the variables output by τ-*GREEDY *in rounds $1, \ldots, s$, i.e., $I_1 \cup \cdots \cup I_s = \{i_1, \ldots, i_r\}$. Given that*

$$\{x_{i_1}, \ldots, x_{i_r}\} \subseteq \tau\text{-}\operatorname{acc}(f) ,$$

the following statements hold.

(a) Let $I^* \subseteq \mathrm{rel}(f) \backslash \{i_1, \ldots, i_r\}$ such that $|I^*| \le \tau$ and $I^* \cap \tau\text{-}\mathrm{inacc}(f) \neq \emptyset$. Then the random variables $|E_{I^*}^{(s)}|$ and all $|E_J^{(s)}|$ with $\emptyset \subsetneq J \subseteq \mathrm{irrel}(f)$, conditional to any fixed values of $x_{i_1}^k, \ldots, x_{i_r}^k$ and $y^k = f(x^k)$, $k \in [m]$, are independent and identically distributed.

(b) The probability that $I_{s+1} \subseteq \mathrm{rel}(f)$ and $I_{s+1} \cap \tau\text{-}\mathrm{inacc}(f) \neq \emptyset$ is at most

$$\frac{d \cdot V(d-1, \tau-1)}{V(n-d, \tau)} \, . \tag{4.9}$$

Proof. (a) Let $R = \{i_1, \ldots, i_r\}$. We show that for $a, b \in \{0,1\}$ and $v \in \{0,1\}^R$ such that there exists $x \in \{0,1\}^n$ with $x|_R = v$ and $f(x) = b$,

$$\Pr\left[x_{I^*} = a \;\middle|\; x|_R = v \wedge f(x) = b\right] = \tfrac{1}{2} \, , \tag{4.10}$$

just as for the variables x_J with $\emptyset \subsetneq J \subseteq \mathrm{irrel}(f)$. Let $I \subseteq R$. If $I \subseteq \tau\text{-}\mathrm{acc}(f)$, then since $|I^*| \le \tau$ and since I^* contains some variable that is not τ-accessible, $\hat{f}(I \cup I^*) = 0$. Otherwise, I contains some irrelevant variable index, and hence $\hat{f}(I \cup I^*) = 0$ by Lemma 2.3.4. By Corollary 2.2.5, $\widehat{f_v}(I^*) = 0$ for all $v \in \{0,1\}^R$. Now the remainder of the proof of (4.10) is identical to the corresponding part of the proof of Lemma 4.3.3, except that I^* has to be substituted for i and r has to be substituted for s. In addition, the reasoning that $|E_{I^*}^{(s)}|$ and all $|E_J^{(s)}|$, $\emptyset \subsetneq J \subseteq \mathrm{irrel}(f)$, are identically distributed follows the same line as the proof of the corresponding statement in Lemma 4.3.3. The independence of $|E_{I^*}^{(s)}|$ and $|E_J^{(s)}|$ is clear by the disjointness of I^* and J. For nonempty $J, J' \subseteq \mathrm{irrel}(f)$ with $J \neq J'$, the independence of $|E_J^{(s)}|$ and $|E_{J'}^{(s)}|$ follows from the independence of the outcomes $\chi_J(x)$, $\chi_{J'}(x)$, and $f(x)$, $x \sim U_n$.

(b) Let $x_{i^*} \in \mathrm{rel}(f) \cap \tau\text{-}\mathrm{inacc}(f)$. The probability that $x_{i^*} \in I_{s+1}$ is equal to the probability that there exists some $I^* \subseteq \mathrm{rel}(f) \setminus \{i_1, \ldots, i_r\}$ with $i^* \in I^*$ that is selected in round $s+1$. By part (a), $|E_{I^*}^{(s)}|$ and $|E_J^{(s)}|$ are identically distributed for all nonempty $J \subseteq \mathrm{irrel}(f)$. Since there are $V(n-d, \tau) - 1$ such sets J with $|J| \le \tau$, I^* is selected in round $s+1$ with probability at most $1/V(n-d, \tau)$. The number of sets $I^* \subseteq \mathrm{rel}(f)$ containing i^* is equal to $V(d-1, \tau-1)$, so the probability that i^* is among the selected variables in round $s+1$ is at most $V(d-1, \tau-1)/V(n-d, \tau)$.

Since there are at most d variables in $\mathrm{rel}(f) \cap \tau\text{-}\mathrm{inacc}(f)$, the probability that I_{s+1} contains at least one out of these variables is at most as large as claimed in (4.9). □

To avoid too complicated expressions, we confine ourselves to present the main negative result for τ-Greedy for simple failure only (instead of λ-failure as in Theorem 4.3.4).

Theorem 4.4.4. *Let $f : \{0,1\}^n \to \{0,1\}$ be a concept that is not τ-Fourier-accessible for some $\tau \in [n]$, and let $d = |\operatorname{rel}(f)|$. Given a uniformly distributed sample S for f of arbitrary size, τ-GREEDY fails on input S with probability at least*

$$1 - \frac{d^2 \cdot V(d-1, \tau-1)}{V(n-d, \tau)} ,$$

where $d = |\operatorname{rel}(f)|$.

Proof. If τ-GREEDY succeeds on input S, then in some round s, τ-GREEDY has to select a set $I_s \subseteq \operatorname{rel}(f)$ with $I_s \cap \tau\text{-}\operatorname{inacc}(f) \neq \emptyset$. The probability that this happens in round s is at most $d \cdot V(d-1, \tau-1)/V(n-d, \tau)$ by Lemma 4.4.3. Since the algorithm can run for at most d rounds if it is successful, the claim follows. □

Plugging $\tau = 1$ into Theorem 4.4.4, we recover Theorem 4.3.4 for the case $\lambda = 1$.

Corollary 4.4.5. *Let $p^{(\tau)}(n, d)$ denote the probability that for any given concept $f : \{0,1\}^n \to \{0,1\}$ with $|\operatorname{rel}(f)| = d$ that is* not *τ-Fourier-accessible and any uniformly distributed sample S for f, τ-GREEDY fails. Then*

(a) for fixed d, $\lim_{n\to\infty} p^{(\tau)}(n, d) = 1$ and

(b) for $d \to \infty$ and $n = n(d) \in \omega(d^2)$, $\lim_{d\to\infty} p^{(\tau)}(n, d) = 1$.

In both items, τ may vary arbitrarily with growing n and/or d. More precisely, in (a), $\tau = \tau(n)$ may be any function $\tau : \mathbb{N} \to [d]$, while in (b), $\tau = \tau(d)$ may be any function $\tau : \mathbb{N} \to \mathbb{N}$ with $\tau(d) \in [d]$ for all $d \in \mathbb{N}$.

4.5 Robustness against Noise

In this section, we investigate the vulnerability of GREEDY when the data is exposed to noise. To our surprise, it turns out that GREEDY is *extremely robust* with respect to heavy noise. More precisely, let P be a γ_a-bounded attribute noise distribution ($\gamma_a > 0$) and $\eta \neq 1/2$ be a classification noise rate (see Section 3.2 for definitions). Given a (P, η)-noisy sample S for a Fourier-accessible concept f of size polynomial in 2^d, $\log(n/\delta)$, γ_a^{-d}, and γ_b, GREEDY still outputs all relevant variables of f, where $\gamma_b = |1 - 2\eta| > 0$.

However, it may be the case that the sets E_i that correspond to the relevant variables do not suffice to explain all edges in E due to noise. Even worse, it may happen that some edges cannot be explained at all: the sample may contain contradictive examples. For this reason, we introduce a variant of GREEDY

Algorithm 4.3 GREEDY$_d$

```
input $S = ((x_1^k, \ldots, x_n^k), y^k)_{k \in [m]}$
$E \leftarrow \{\{k, \ell\} \mid k, \ell \in [m], y^k \neq y^\ell\}$
$R \leftarrow \emptyset$
while $|R| < d$ do
  for $i = 1$ to $n$ do
    $E_i \leftarrow \{\{k, \ell\} \in E \mid x_i^k \neq x_i^\ell\}$
  select $x_i \notin R$ with maximum $|E_i|$
  $E \leftarrow E \setminus E_i$
  $R \leftarrow R \cup \{x_i\}$
output GREEDY$_d(S) = R$
```

that is given the number d of relevant attributes as a parameter and outputs exactly d variables. We denote this algorithm by GREEDY$_d$. It is presented as Algorithm 4.3.

We adjust definition (4.2) of the probabilities α_I^{ab} and define for $a, b \in \{0, 1\}$

$$\beta_I^{ab} = \Pr[x_I \oplus \xi_I = a \wedge f(x) \oplus \zeta = b] \ , \tag{4.11}$$

where $\xi \sim P$, $\Pr[\zeta = 1] = \eta$, and $\Pr[\zeta = 0] = 1 - \eta$.

While in the noise-free scenario, the expectation of $|E_I|$ is $\alpha_I(m-1)m$, the expectation of $|E_I|$ is now equal to $\beta_I(m-1)m$ with

$$\beta_I = \beta_I^{00}\beta_I^{11} + \beta_I^{10}\beta_I^{01} \ . \tag{4.12}$$

The next lemma is completely analogous to Lemma 4.2.2:

Lemma 4.5.1. *There exist $c_1, c_2 > 0$ such that for all $f : \{0,1\}^n \to \{0,1\}$, given a uniformly distributed (P, η)-noisy sample S of size m for f, for all $I \subseteq [n]$ and arbitrary ϵ with $0 \leq \epsilon \leq 1$,*

$$\Pr\left[\left||E_I| - \beta_I m^2\right| > \epsilon m^2\right] \ < \ c_1 e^{-c_2 \epsilon^2 m} \ .$$

Proof. The proof is identical to that of Lemma 4.2.2, except that every α has to be replaced with β and the sets A_I^{ab}, $a, b \in \{0, 1\}$, have to be replaced with sets

$$B_I^{ab} = \{k \in [m] \mid (x_I^k \oplus \xi_I^k, y^k \oplus \zeta^k) = (a, b)\} \ .$$

□

Proving an analog of Lemma 4.2.3 requires some more computation:

Lemma 4.5.2. *Let $f : \{0,1\}^n \to \{0,1\}$ and $I \subseteq [n]$ with $I \neq \emptyset$. Let P be an attribute noise distribution and η be a classification noise rate. Then*

$$\beta_I = \tfrac{1}{2}\Big((1-2\eta)^2 \operatorname{Var}[f] + \eta(1-\eta) + (1-2\eta)^2(1-2p_I)^2 \hat{f}(I)^2\Big) \tag{4.13}$$

with $p_I = \Pr_{\xi\sim P}[\xi_I = 1]$ (as defined in (3.4)).

Proof. We first express β_I^{ab} in terms of α_I^{ab}, p_I, and η. According to (4.11), we have

$$\begin{aligned}
\beta_I^{ab} &= \Pr[x_I = a \wedge f(x) = b \wedge \xi_I = 0 \wedge \zeta = 0] \\
&\quad + \Pr[x_I = a \wedge f(x) = 1-b \wedge \xi_I = 0 \wedge \zeta = 1] \\
&\quad + \Pr[x_I = 1-a \wedge f(x) = b \wedge \xi_I = 1 \wedge \zeta = 0] \\
&\quad + \Pr[x_I = 1-a \wedge f(x) = 1-b \wedge \xi_I = 1 \wedge \zeta = 1] \\
&= \alpha_I^{ab}(1-p_I)(1-\eta) + \alpha_I^{a,1-b}(1-p_I)\eta \\
&\quad + \alpha_I^{1-a,b} p_I(1-\eta) + \alpha_I^{1-a,1-b} p_I \eta\ .
\end{aligned} \tag{4.14}$$

Since $\alpha_I^{a,1-b} = \frac{1}{2} - \alpha_I^{a,b}$, we further obtain

$$\begin{aligned}
\beta_I^{ab} &= \alpha_I^{ab}\left[(1-p_I)(1-\eta) - (1-p_I)\eta\right] + \tfrac{1}{2}(1-p_I)\eta \\
&\quad + \alpha_I^{1-a,b}\left[p_I(1-\eta) - p_I\eta\right] + \tfrac{1}{2}p_I\eta \\
&= \alpha_I^{ab}(1-p_I)(1-2\eta) + \alpha_I^{1-a,b} p_I(1-2\eta) + \tfrac{1}{2}\eta\ .
\end{aligned}$$

Now we plug this into (4.12) and obtain

$$\begin{aligned}
\beta_I &= \left(\alpha_I^{00}(1-p_I)(1-2\eta) + \alpha_I^{10}p_I(1-2\eta) + \tfrac{1}{2}\eta\right) \\
&\qquad \cdot \left(\alpha_I^{11}(1-p_I)(1-2\eta) + \alpha_I^{01}p_I(1-2\eta) + \tfrac{1}{2}\eta\right) \\
&\quad + \left(\alpha_I^{10}(1-p_I)(1-2\eta) + \alpha_I^{00}p_I(1-2\eta) + \tfrac{1}{2}\eta\right) \\
&\qquad \cdot \left(\alpha_I^{01}(1-p_I)(1-2\eta) + \alpha_I^{11}p_I(1-2\eta) + \tfrac{1}{2}\eta\right) \\
&= \alpha_I^{00}\alpha_I^{11}\left[(1-p_I)^2(1-2\eta)^2 + p_I^2(1-2\eta)^2\right] \\
&\quad + \alpha_I^{00}\alpha_I^{01}\left[p_I(1-p_I)(1-2\eta)^2 + p_I(1-p_I)(1-2\eta)^2\right] \\
&\quad + \alpha_I^{10}\alpha_I^{11}\left[p_I(1-p_I)(1-2\eta)^2 + p_I(1-p_I)(1-2\eta)^2\right] \\
&\quad + \alpha_I^{10}\alpha_I^{01}\left[p_I^2(1-2\eta)^2 + (1-p_I)^2(1-2\eta)^2\right] \\
&\quad + \tfrac{1}{2}\eta(1-2\eta)\left(\alpha_I^{00} + \alpha_I^{11} + \alpha_I^{10} + \alpha_I^{01}\right) + \tfrac{1}{2}\eta^2 \\
&= \alpha_I\left(p_I^2 + (1-p_I)^2\right)(1-2\eta)^2 \\
&\quad + \left(\alpha_I^{00}\alpha_I^{01} + \alpha_I^{10}\alpha_I^{11}\right)\cdot 2 \cdot p_I(1-p_I)(1-2\eta)^2 \\
&\quad + \tfrac{1}{2}\eta(1-2\eta) + \tfrac{1}{2}\eta^2\ .
\end{aligned}$$

We obtain an easier expression for $\alpha_I^{00}\alpha_I^{01} + \alpha_I^{10}\alpha_I^{11}$ using (4.6) and the equation $\alpha_I^{01} - \alpha_I^{11} = \alpha_I^{10} - \alpha_I^{00}$, which follows from $\alpha_I^{00} + \alpha_I^{01} = \frac{1}{2} = \alpha_I^{10} + \alpha_I^{11}$:

$$\begin{aligned}\alpha_I - \hat{f}(I)^2 &= \alpha_I^{00}\alpha_I^{11} + \alpha_I^{10}\alpha_I^{01} - (\alpha_I^{01} - \alpha_I^{11})^2 \\ &= \alpha_I^{00}\alpha_I^{11} + \alpha_I^{10}\alpha_I^{01} - (\alpha_I^{01} - \alpha_I^{11})(\alpha_I^{10} - \alpha_I^{00}) \\ &= \alpha_I^{00}\alpha_I^{01} + \alpha_I^{10}\alpha_I^{11} \ .\end{aligned}$$

Plugging this into the above formula, we obtain

$$\begin{aligned}\beta_I &= \alpha_I \left[(p_I^2 + (1-p_I)^2)(1-2\eta)^2 + 2p_I(1-p_I)(1-2\eta)^2\right] \\ &\quad -2\hat{f}(I)^2 p_I(1-p_I)(1-2\eta)^2 + \tfrac{1}{2}\eta(1-2\eta) + \tfrac{1}{2}\eta^2 \\ &= \alpha_I(1-2\eta)^2 - 2\hat{f}(I)^2 p_I(1-p_I)(1-2\eta)^2 + \tfrac{1}{2}\eta(1-\eta) \ .\end{aligned}$$

Finally, (4.5) yields

$$\beta_I = \tfrac{1}{2}\operatorname{Var}[f](1-2\eta)^2 + \hat{f}(I)^2\left(\tfrac{1}{2} - 2p_I(1-p_I)\right)(1-2\eta)^2 + \tfrac{1}{2}\eta(1-\eta) \ .$$

Since $\frac{1}{2} - 2p_I(1-p_I) = \frac{1}{2}(1-2p_I)^2$, the claim follows. □

Theorem 4.3.1 generalizes to the scenario of noisy data as follows:

Theorem 4.5.3. *There is a polynomial p such that the following holds. Let $f : \{0,1\}^n \to \{0,1\}$ be a Fourier-accessible concept, $d = |\operatorname{rel}(f)|$, P be a γ_a-bounded attribute noise distribution, η be a classification noise rate satisfying $|1-2\eta| \geq \gamma_b > 0$, and $\delta > 0$. Let S be a uniformly distributed (P,η)-noisy sample S for f of size*

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

Then $\textsc{Greedy}_d(S) = \operatorname{rel}(f)$ *with probability at least* $1-\delta$.

Proof. Let $x_{i_1}, \ldots, x_{i_d}$ be the sequence of variables output by $\textsc{Greedy}_d$. For $s \in [d]$, let $R_s = \{i_1, \ldots, i_s\}$. The proof is similar to the proof of Theorem 4.3.1, so we only point out the differences. First, ϵ has to be chosen as $2^{-3d-3}\gamma_a^{2d}\gamma_b^2$. For each $I \subseteq [n]$ with $1 \leq |I| \leq d$, we have

$$\Pr\left[\left||E_I| - \beta_I m^2\right| > \epsilon m^2\right] < c_1 e^{-c_2\epsilon^2 m}$$

for some constants $c_1, c_2 > 0$ by Lemma 4.5.1. Consequently,

$$\forall I \subseteq [n] \text{ such that } 1 \leq |I| \leq d : \ \left||E_I| - \beta_I m^2\right| \leq \epsilon m^2 \tag{4.15}$$

with probability at least $\rho = 1 - n^d \cdot c_1 e^{-c_2\epsilon^2 m}$ (since $V(n,d) - 1 \leq n^d$). In the following, we assume that (4.15) holds. Thus, all subsequent consequences of (4.15) hold with probability at least ρ.

Next, the calculation of $|E_{i^*}^{(s)}| - |E_j^{(s)}|$ has to be adjusted:

$$\begin{aligned}
|E_{i^*}^{(s)}| - |E_j^{(s)}| &= 2^{-s} \sum_{\emptyset \subseteq I \subseteq R_s} \left(|E_{I\cup\{i^*\}}| - |E_{I\cup\{j\}}|\right) \\
&\geq 2^{-s} \sum_{\emptyset \subseteq I \subseteq R_s} \left((\beta_{I\cup\{i^*\}} - \epsilon)m^2 - (\beta_{I\cup\{j\}} + \epsilon)m^2\right) \\
&\geq 2^{-s} \sum_{\emptyset \subseteq I \subseteq R_s} \left(\tfrac{1}{2}(1-2\eta)^2 \operatorname{Var}[f] + \tfrac{1}{2}\eta(1-\eta)\right. \\
&\qquad\qquad + \tfrac{1}{2}(1-2\eta)^2(1-2p_I)^2 \hat{f}(I)^2 - \epsilon \\
&\qquad\qquad \left. - \tfrac{1}{2}(1-2\eta)^2 \operatorname{Var}[f] - \tfrac{1}{2}\eta(1-\eta) - \epsilon\right) \\
&\geq 2^{-s} \left(\tfrac{1}{2}(1-2\eta)^2(1-2p_I)^2 \hat{f}(I_0 \cup \{i^*\})^2 + 2^s \cdot (-2\epsilon)\right) m^2 \\
&\geq (2^{-3d-1}\gamma_b^2\gamma_a^{2d} - 2\epsilon)m^2 \;\geq\; 2^{-3d-2}\gamma_b^2\gamma_a^{2d} m^2 > 0 \ .
\end{aligned}$$

Finally, it suffices to choose $m \geq m_1 = c_2^{-1} \cdot \gamma_a^{-4d}\gamma_b^{-2} \cdot 2^{6d-6}(d \ln n + \ln(c_1/\delta))$ to have $\rho \geq 1 - \delta$. This is polynomial in 2^d, $\log n$, $\log(1/\delta)$, γ_a^{-d}, and γ_b^{-1}. □

As learning without noise can be considered as a special case of learning under noise, the lower bound stated in Theorem 4.3.4 is also valid for the noisy scenario, in the sense that there exists a noise distribution (namely, with $P(0^n) = 1$ and $\eta = 0$) such that Theorem 4.3.4 holds. So far, however, we cannot exclude that there may be noise distributions for which it turns out that *strictly more* concepts may be learned under non-degenerate noise (in the sense that some data are flipped with positive probability) than can be learned without noise. This counter-intuitive possibility is ruled out in the following by extending also the negative result of Section 4.3.2 to the noisy scenario.

First, we observe that Lemma 4.3.3 also holds if S is a uniformly distributed (P, η)-noisy sample. To see this, one may replace all occurrences of x with $x \oplus \xi$ and all occurrences of $f(x)$ with $f(x) \oplus \zeta$ in the proof of Lemma 4.3.3 (a). The only place where the proof becomes invalid is equation (4.8). However, using (4.14), it is easy to see that

$$\beta_i^{01} - \beta_i^{11} = \beta_i^{10} - \beta_i^{00} = (1-2p_i)(1-2\eta)\hat{f}(i) \ .$$

Since $p_i \neq 1/2$ and $\eta \neq 1/2$, this quantity vanishes if and only if $\hat{f}(i)$ does. Hence, this generalizes the proof for noisy samples.

Now Theorem 4.3.4 and Corollary 4.3.5 also hold for uniformly distributed (P, η)-noisy samples, using exactly the same proofs.

Algorithm 4.4 Greedy Ranking

```
input $S = ((x_1^k, \ldots, x_n^k), y^k)_{k \in [m]}$
$E \leftarrow \{\{k,\ell\} \mid k,\ell \in [m], y^k \neq y^\ell\}$
$R \leftarrow \emptyset$
for $i = 1$ to $n$ do
    $E_i \leftarrow \{\{k,\ell\} \in E \mid x_i^k \neq x_i^\ell\}$
while $E \neq \emptyset$ do
    select $x_i \notin R$ with maximum $|E_i|$
    $E \leftarrow E \setminus E_i$
    $R \leftarrow R \cup \{x_i\}$
output Greedy Ranking$(S) = R$
```

4.6 Greedy Ranking

In this section, we present Greedy Ranking (Algorithm 4.4), an even easier variant of Greedy. Recall that Greedy dynamically recomputes the sets E_i after each round. In contrast, the static variant Greedy Ranking simply ranks the sets E_i by their size only once in the beginning. Then it successively selects the variables corresponding to the largest sets E_i until all edges of the functional relations graph are covered.

Even though this strategy can be arbitrarily bad for the Set Cover problem in terms of its approximation ratio, it turns out that applied to the problem of learning relevant attributes, Greedy Ranking often performs equally well as Greedy.

Since we have already seen in Section 4.2 that the expectation of $|E_i|$ is $\frac{1}{2}(\mathrm{Var}[f] + \hat{f}(i)^2)m(m-1)$, it is not very surprising that Greedy Ranking succeeds with high probability, provided that the target concept f is 1-low (see Definition 2.4.1):

Theorem 4.6.1. *There is a polynomial p such that the following holds. Let $f : \{0,1\}^n \to \{0,1\}$ be a 1-low concept, $d = |\mathrm{rel}(f)|$, and $\delta > 0$. Let S be a uniformly distributed sample for f of size*

$$m \geq p(2^d, \log n, \log(1/\delta)) .$$

Then Greedy Ranking$(S) = \mathrm{rel}(f)$ *with probability at least* $1 - \delta$.

Proof. For the same reasons that were presented in the proof Theorem 4.3.1, Greedy Ranking outputs *at least d* variables with probability at least $1 - \delta$, provided that $m \geq m_0 = 2^{2d+1} \ln(2n/\delta)$.

For each $i \in [n]$, we have

$$\Pr\left[\left||E_i| - \alpha_i m^2\right| > \epsilon m^2\right] < c_1 e^{-c_2 \epsilon^2 m}$$

for some constants $c_1, c_2 > 0$ by Lemma 4.2.2. Consequently,

$$\forall i \in [n] : \ \big||E_i| - \alpha_i m^2\big| \leq \epsilon m^2 \tag{4.16}$$

with probability at least $\rho = 1 - n \cdot c_1 e^{-c_2 \epsilon^2 m}$.

Let $\epsilon = 2^{-2d-2}$. Let $x_i \in \mathrm{rel}(f)$ and $x_j \in \mathrm{irrel}(f)$. Then $\hat{f}(j) = 0$ (by Lemma 2.3.4) and $|\hat{f}(i)| \geq 2^{-d}$ (by Lemma 2.3.6). Assuming that (4.16) holds, we have

$$\begin{aligned} |E_i| - |E_j| &\geq (\alpha_i - \epsilon)m^2 - (\alpha_j + \epsilon)m^2 \\ &= \tfrac{1}{2}\Big(\mathrm{Var}[f] + \hat{f}(i)^2 - \epsilon\Big) m^2 - \tfrac{1}{2}\Big(\mathrm{Var}[f] + \hat{f}(j)^2 + \epsilon\Big) m^2 \\ &\geq \tfrac{1}{2}\hat{f}(i)^2 - \epsilon \ \geq \ 2^{-2d-1} - \epsilon \ > \ 0 \ . \end{aligned}$$

Consequently, each set $|E_i|$, $x_i \in \mathrm{rel}(f)$, is larger than all sets $|E_j|$, $x_j \in \mathrm{irrel}(f)$. Hence, GREEDY RANKING selects all relevant variables and then halts.

It suffices to choose $m \geq m_1 = c_2^{-1} \cdot 2^{4d-2}(\ln n + \ln(2c_1/\delta))$ to have $\rho \geq 1 - \delta/2$. In total, we can choose $m = \max\{m_0, m_1\}$ to guarantee that GREEDY RANKING outputs exactly the relevant variables of f. This amount is polynomial in 2^d, $\log n$, and $\log(1/\delta)$. □

A characterization of 1-lowness is provided by Lemma 2.4.2, examples of 1-low functions are given in Example 2.4.4. Specifically, a concept with symmetric base function is 1-low if and only if it is Fourier-accessible. Hence, GREEDY and GREEDY RANKING perform equally well for symmetric target concepts.

Furthermore, having in mind the proof of Theorem 4.3.4 which states that GREEDY fails for all concepts that are not Fourier-accessible, it is not hard to see that GREEDY RANKING fails for concepts that are *not* 1-low.

Theorem 4.6.2. *Let $f : \{0,1\}^n \to \{0,1\}$ be a concept that is not 1-low and $\lambda \geq 1$. Given a uniformly distributed sample S for f of arbitrary size,* GREEDY RANKING *λ-fails on input S with probability at least*

$$1 - \frac{\lambda d^2}{n - \lambda d} ,$$

where $d = |\,\mathrm{rel}(f)|$.

Proof. Let x_i be a variable that is not 1-low for f. Applying Lemma 4.3.3 (a) with $s = 0$ yields that $|E_i| = |E_i^{(0)}|$ and $|E_j| = |E_j^{(0)}|$, $x_j \in \mathrm{irrel}(f)$, are independent and identically distributed. Thus, in each round r of the `while`-loop, conditional to having output only relevant variables in rounds $1, \ldots, r-1$, the probability that GREEDY RANKING outputs x_i is at most $1/(n - r + 1)$. Having at most d non-1-low variables and $\lambda \cdot d$ rounds available, GREEDY RANKING succeeds with probability at most $\frac{\lambda \cdot d^2}{n - \lambda \cdot d}$. □

Algorithm 4.5 τ-Greedy Ranking.

```
input $S = ((x_1^k, \dots, x_n^k), y^k)_{k \in [m]}$
$E \leftarrow \{\{k, \ell\} \mid k, \ell \in [m], y^k \neq y^\ell\}$
$R \leftarrow \emptyset$
for $I \subseteq [n]$ with $1 \leq |I| \leq \tau$ do
    $E_I \leftarrow \{\{k, \ell\} \in E \mid x_I^k \neq x_I^\ell\}$
while $E \neq \emptyset$ do
    select $I \subseteq [n]$, $1 \leq |I| \leq \tau$, with maximum $|E_I|$
    $E \leftarrow E \setminus \bigcup_{i \in I} E_i$
    $R \leftarrow R \cup \{x_i \mid i \in I\}$
output $\tau$-Greedy Ranking$(S) = R$
```

Also the extension of Greedy Ranking to τ-Greedy Ranking (Algorithm 4.5) that ranks all sets $|E_I|$, $1 \leq |I| \leq \tau$, follows the same line as the extension of Greedy to τ-Greedy.

Theorem 4.6.3. *There is a polynomial p such that the following holds. Let $f : \{0,1\}^n \to \{0,1\}$ be a τ-low concept, $1 \leq \tau \leq d = |\operatorname{rel}(f)|$, and $\delta > 0$. Let S be a uniformly distributed sample for f of size*

$$m \geq p(2^d, \log n, \log(1/\delta)) .$$

*Then τ-*Greedy Ranking*$(S) = \operatorname{rel}(f)$ with probability at least $1 - \delta$.*

Proof. The proof is very similar to the proof of Theorem 4.6.1 and a detailed description therefore omitted; see also the proof of Theorem 4.4.2, which is the extension of Theorem 4.3.1 to τ-Fourier-accessible concepts. □

Also the lower bound for τ-Greedy Ranking and its proof are almost identical to the proof of the corresponding results for τ-Greedy, Theorem 4.4.4 (with the aid of Lemma 4.4.3) and Corollary 4.4.5. Thus, we leave out detailed descriptions of the proofs of the following statements.

Theorem 4.6.4. *Let $f : \{0,1\}^n \to \{0,1\}$ be a concept that is not τ-low for some $\tau \in [n]$. Given a sample S for f of arbitrary size, τ-*Greedy Ranking *fails on input S with probability at least*

$$1 - \frac{d^2 \cdot V(d-1, \tau-1)}{V(n-d, \tau)} ,$$

where $d = |\operatorname{rel}(f)|$.

Algorithm 4.6 Greedy Ranking$_d$

```
input $S = ((x_1^k, \dots, x_n^k), y^k)_{k \in [m]}$
$E \leftarrow \{\{k, \ell\} \mid k, \ell \in [m], y^k \neq y^\ell\}$
$R \leftarrow \emptyset$
for $i = 1$ to $n$ do
    $E_i \leftarrow \{\{k, \ell\} \in E \mid x_i^k \neq x_i^\ell\}$
while $|R| < d$ do
    select $x_i \notin R$ with maximum $|E_i|$
    $E \leftarrow E \setminus E_i$
    $R \leftarrow R \cup \{x_i\}$
output Greedy Ranking$_d(S) = R$
```

Corollary 4.6.5. *Let $p^{(\tau)}(n,d)$ denote the probability that for any given concept $f : \{0,1\}^n \to \{0,1\}$ with $|\operatorname{rel}(f)| = d$ that is* not *τ-Fourier-accessible and any uniformly distributed sample S for f, τ-*Greedy Ranking *fails. Then*

(a) for fixed d, $\lim_{n\to\infty} p^{(\tau)}(n,d) = 1$ and

(b) for $d \to \infty$ and $n = n(d) \in \omega(d^2)$, $\lim_{d\to\infty} p^{(\tau)}(n,d) = 1$.

In both items, τ may vary arbitrarily with growing n and/or d. More precisely, in (a), $\tau = \tau(n)$ may be any function $\tau : \mathbb{N} \to [d]$, while in (b), $\tau = \tau(d)$ may be any function $\tau : \mathbb{N} \to \mathbb{N}$ with $\tau(d) \in [d]$ for all $d \in \mathbb{N}$.

If noise is introduced, then—as for Greedy—also Greedy Ranking has to be modified by providing the number of relevant attributes as an additional parameter, resulting in Greedy Ranking$_d$, which is presented as Algorithm 4.6.

Theorem 4.6.1 can be extended to the scenario of noisy data in the same way as we have extended Theorem 4.3.1 to Theorem 4.5.3:

Theorem 4.6.6. *There is a polynomial p such that the following holds. Let $f : \{0,1\}^n \to \{0,1\}$ be a 1-low concept, $d = |\operatorname{rel}(f)|$, P be a γ_a-bounded attribute noise distribution, η be a classification noise rate with $|1 - 2\eta| \geq \gamma_b > 0$, and $\delta > 0$. Let S be a uniformly distributed (P,η)-noisy sample for f of size*

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

Then Greedy Ranking$_d(S) = \operatorname{rel}(f)$ *with probability at least $1 - \delta$.*

Also the lower bound can be generalized to the noisy scenario for Greedy Ranking, i.e., Theorem 4.6.2 is valid for (P,η)-noisy samples as well. We omit the proof.

4.7 Modest Ranking —the Opposite of Greedy Ranking

We have shown in Subsection 4.3.3 that the characterization we have found for the setting of uniformly distributed attributes cannot be generalized to non-uniformly distributed attributes. In the latter case, it can even happen that the opposite strategy to greedy is applicable: here, attributes that correspond to *smallest* set cardinalities are selected first. Such an algorithm works correctly with high probability if the sets E_i for relevant attributes are likely to be *smaller* than the sets E_i for the irrelevant attributes. Instead of being greedy, this algorithm rather behaves modestly, so we call it Modest Ranking. The pseudocode for Modest Ranking is the same as for Greedy Ranking (Algorithm 4.4), except that in line 7, a set of *minimum* size is selected.

For uniformly distributed examples, this strategy is of no use since the sets E_i for $x_i \in \mathrm{rel}(f)$ are always at least as large as the sets E_j for $x_j \in \mathrm{irrel}(f)$ (in expectation).

In terms of their ability to find small set covers, none of the strategies Greedy, Greedy Ranking, and Modest Ranking outperforms the other two on all instances. In other words, for each of the strategies, it is possible to construct (even small) set cover instances such that this strategy finds a minimal cover, whereas the other two strategies fail to do so.

To see that Modest Ranking is indeed applicable in some situations, consider the following scenario. Let $f\colon \{0,1\}^n \to \{0,1\}$ with base function $f'\colon \{0,1\}^d \to \{0,1\}$ defined by $f(x_1,\dots,x_d) = x_1 \wedge \cdots \wedge x_d$, let $d^* \in [0,1]$, and let D be the product distribution for which all rates are equal to d^*. We have shown [AR03, Theorem 4] that if $d^* \leq 1/2$, then Greedy Ranking is successful with high probability (the case $d^* = 1/2$ is also covered by Theorem 4.6.1) and that if $d^* > 1/2$, then Modest Ranking is successful with high probability. The same ideas apply to the OR-function, except that we have to substitute $1 - d^*$ for d^*.

CHAPTER 5

The Fourier Method

In this chapter, we devise Fourier-based algorithms that learn the class of τ-low concepts in the presence of quite general attribute and classification noise. In particular, we show how to produce accurate hypotheses (instead of merely learning the relevant attributes). As an application, the class of monotone juntas can be learned efficiently in this setting. We generalize the Fourier approach to non-uniformly distributed attributes and show that the relevant attributes of monotone juntas and of parity juntas can be learned efficiently in the general noise model under consideration. It turns out that the construction of a suitable hypothesis is a lot more intricate in this situation than it is for uniformly distributed attributes. Nevertheless, we do obtain some positive results.

After reviewing how to learn juntas via the Fourier method in the noise-free case in Section 5.1, we show in Section 5.2 how the noisy case can be handled as an application of known results, using a sample size of roughly n^d. In Section 5.3, we show how to learn the relevant attributes from a sample size that grows only logarithmically in n. The overall learning algorithm including the construction of the hypothesis is described in Section 5.4. We extend our results to non-uniformly distributed attributes in Sections 5.5, 5.6, and 5.7.

In this chapter, all concepts map to the range $\Omega = \{-1, +1\}$. As a consequence, classification noise bits ζ are also supposed to be elements of $\{-1, +1\}$, drawn with $\Pr[\zeta = -1] = \eta$ and $\Pr[\zeta = +1] = 1 - \eta$ for some classification noise rate $\eta \in [0, 1]$, which we indicate by $\zeta \sim \eta$. The classification y of a random example $(x, y) \in \{0, 1\}^n \times \{-1, +1\}$ is then *multiplied* by ζ. Thus, each classification is independently affected by noise with probability η.

Algorithm 5.1 τ-FOURIER$_d$

```
input $S = ((x_1^k, \ldots, x_n^k), y^k)_{k \in [m]}$
$R \leftarrow \emptyset$
for $I \subseteq [n]$ with $1 \leq |I| \leq \tau$ 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 $\tau$-FOURIER$_d(S) = R$
```

5.1 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. [MOS04]. 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* (which we have introduced in Definition 2.4.1) and to understand the methods presented in Section 5.2.

Let $f : \{0,1\}^n \rightarrow \{-1,+1\}$ be a monotone d-junta. Then f is correlated with all of its relevant variables [MOS04], i.e., $\hat{f}(i) \neq 0$ for all $x_i \in \mathrm{rel}(f)$. 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 (3.2). 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., $|\hat{f}(i)| \geq \rho$ for some threshold $\rho > 0$.

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 Lemma 3.5.1), 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 concepts with the property that all relevant variables are correlated with the function value, i.e., the 1-low concepts. Moreover, we can use Lemma 2.3.4 to extend the method to larger classes of Boolean concepts by looking beyond the first level of Fourier coefficients.

The algorithm (which we call τ-FOURIER$_d$) for inferring the relevant variables of τ-low d-juntas, which has been described by Mossel et al. [MOS04], is presented as Algorithm 5.1.

Theorem 5.1.1 ([MOS04]). *There exist polynomials p_s and p_t such that the following holds. Let $f : \{0,1\}^n \rightarrow \{-1,+1\}$ be a τ-low d-junta, $\delta > 0$, and*

*S be a uniformly distributed sample of size $m \geq p_s(\log n, 2^d, \log(1/\delta))$. Then on input S, τ-*FOURIER$_d$ *outputs exactly the relevant variables of f in time $n^\tau \cdot p_t(n, 2^d, \log(1/\delta))$ with probability at least $1 - \delta$.*

5.2 The Noisy Case: Uniform Attribute Distribution

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 $\eta \in [0,1]$. As we have already discussed in Section 3.2, it is reasonable to assume that $\eta \neq 1/2$. Thus, we assume that there exists some bound $\gamma_b > 0$ such that $|1 - 2\eta| \geq \gamma_b$.

Recall the definition $p_I = \Pr_{\xi \sim P}[\chi_I(\xi) = -1]$. As we will approximate Fourier coefficients from noisy examples, it is also necessary to require the probabilities p_I to be different from $1/2$ for all $I \subseteq [n]$ with $|I| \leq d$, see Lemma 3.3.2. Moreover, such a restriction is even necessary for an information-theoretic reason: we have seen in Theorem 3.7.3 that $p_I = 1/2$ may cause the concept class to be impossible to learn under attribute noise distribution P. In the following, we will require P to be γ_a-bounded for some fixed $\gamma_a > 0$ (see Definition 3.2.6) since we believe that this is the most important case. Recall that γ_a-bounded attribute noise distributions P include product distributions with rates $p_1, \ldots, p_n$ that satisfy $|1-2p_i| \geq \gamma_a$ for all $i \in [n]$. Extending our results to more general distributions P satisfying $p_I \neq 1/2$ is straightforward: all statements and proofs remain valid if we substitute

$$\lambda_t = \min\{|\lambda_I| \mid I \subseteq [n], |I| \leq t\}$$

for γ_a^t in all places.

Furthermore, we fix a confidence parameter $\delta \in (0,1]$, an accuracy parameter $\epsilon \in (0,1]$, 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.

Since d-juntas have all of their Fourier weight located in levels $0, \ldots, d$ (by Lemma 2.3.4), we obtain a first result for learning the class $\mathcal{J}_d^n$ of n-ary d-juntas from noisy examples by applying NOISY-LMN$_\mathcal{T}(x)$ (Algorithm 3.1) with $\mathcal{T} = \{I \subseteq [n] \mid |I| \leq d\}$.

Theorem 5.2.1. *The class $\mathcal{J}_d^n$ is exactly learnable with confidence $1 - \delta$ from uniformly distributed (P,η)-noisy samples, using sample complexity and running time*

$$n^d \cdot \mathrm{poly}\left(n, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1}\right) .$$

Algorithm 5.2 τ-NOISY-FOURIER$_d$.

```
1: input $S = ((x_1^k, \ldots, x_n^k), y^k)_{k \in [m]}, \gamma_a, \gamma_b$
2: $R \leftarrow \emptyset$
3: for $I \subseteq [n]$ with $1 \leq |I| \leq \tau$ do
4:     $\beta \leftarrow (\gamma_a^{|I|} \cdot \gamma_b)^{-1} \cdot \frac{1}{m} \cdot \sum_{k=1}^{m} \chi_I(x^k) \cdot y^k$
5:     if $|\beta| \geq 2^{-d-1}$
6:         then $R \leftarrow R \cup \{x_i \mid i \in I\}$
7: output $\tau$-NOISY-FOURIER$_d(S) = R$
```

Proof. We choose $\epsilon = 2^{-d-1}$ and $\mathcal{T}_\epsilon = \{I \subseteq [n] \mid |I| \leq d\}$ (since $\hat{f}(I) = 0$ for all I of size larger than d) and apply Theorem 3.6.3. By assumption,

$$\lambda = \min\{|\lambda_I| \mid I \in \mathcal{T}_\epsilon\} \geq \gamma_a^d \ ,$$

and the claim follows. □

Unfortunately, with this approach, 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. [O'D03, Section 3.3]).

In the sequel, we show how to combine the method just described with the idea of first detecting the relevant attributes, as we did in the noise-free case. In Theorem 5.4.1, 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 τ-low d-juntas with $\tau < d$, also the running time decreases from $O(n^{d+O(1)})$ to $O(n^{\tau+O(1)})$.

5.3 Learning Relevant Attributes from Uniformly Distributed Examples

The detection of relevant variables works similarly as in the noise-free case. The following modifications to τ-FOURIER$_d$ (Algorithm 5.1) vaccinate it against noise; the resulting algorithm τ-NOISY-FOURIER$_d$ is presented as Algorithm 5.2.

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$_d$, $x^k = x'^k \oplus \xi^k$ and $y^k = y'^k \cdot \zeta^k$ for appropriate noise-free data x'^k, y'^k and noise ξ^k, ζ^k.

Theorem 5.3.1. *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 3.3.2, for every $I \subseteq [n]$ of size at most τ,

$$|\tilde{f}_S(I) - (1 - 2p_I)(1 - 2\eta)\hat{f}(I)| \leq \tfrac{1}{2} \cdot \gamma_a^\tau \cdot \gamma_b \cdot \rho \tag{5.1}$$

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 $\hat{f}(I) \neq 0$. By Lemma 2.3.6, $|\hat{f}(I)| \geq 2^{-d}$. In particular, if (5.1) 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 $1 - \delta/n$.

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 2.3.4. By (5.1), 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 $n \cdot (\delta/n) = \delta$. □

5.4 Constructing a Hypothesis from Uniformly Distributed Examples

Learning juntas (in the sense of constructing an accurate hypothesis) from noisy data 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 to the algorithm used in the noise-free setting is that we cannot simply read off the truth table from the examples since

Algorithm 5.3 τ-Noisy-Learn$_d$

```
input $S = ((x_1^k, \ldots, x_n^k), y^k)_{k \in [m]}, P, \gamma_a, \eta$
$\gamma_b \leftarrow |1 - 2\eta|$
$R \leftarrow \tau\text{-Noisy-Fourier}_d(S, \gamma_a, \gamma_b)$
$\mathcal{T} \leftarrow \mathcal{P}(R)$
output hypothesis
    $\tau\text{-Noisy-Learn}_d(x) = \text{Noisy-LMN}_{\mathcal{T}}(S, P, \eta)(x)$
```

these may contain inconsistencies (even if not, such a truth table is unlikely to be correct).

Fortunately, we have seen in Section 3.6 how to build a good hypothesis in the presence of attribute noise. The trick is that we do not apply Theorem 3.7.4 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 Theorem 3.7.4). Moreover, for the class $\mathcal{J}_d^n(\tau)$ of τ-low d-juntas, the factor of n^d in the running time reduces to n^τ. Precisely, the algorithm, which we call τ-Noisy-Learn$_d$, is presented as Algorithm 5.3.

Theorem 5.4.1. *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* $\text{poly}(\log n, 2^d, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$
- *with running time* $n^\tau \cdot \text{poly}(n, 2^d, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$.

Proof. Let $f \in \mathcal{J}_d^n(\tau)$. As we have shown in Theorem 5.3.1, 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 Theorem 3.6.3, choosing $\epsilon = 2^{-d-1}$ and $\mathcal{T}_\epsilon = \mathcal{P}(\text{rel}(f))$, again with probability at least $1 - \delta/2$, the output hypothesis exactly coincides with f, provided that $m \geq \text{poly}(|\mathcal{T}_\epsilon|, 1/\lambda, 1/\epsilon, 1/|1 - 2\eta|)$ with λ as defined in (3.14). Since $|\mathcal{T}_\epsilon| \leq 2^d$, $\lambda \leq \gamma_a^{-d}$, and $\epsilon^{-1} = 2^{d+1}$, τ-Noisy-Learn$_d$ succeeds in exactly learning the target concept with probability at least $1 - \delta$, provided that the sample size is $p(\log n, 2^d, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1})$ for some sufficiently large polynomial p. The claimed running time follows from Theorem 5.3.1 and Theorem 3.6.3. □

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

Corollary 5.4.2. *(a) The class* $\mathcal{J}_d^n$ *can be exactly learned with confidence* $1 - \delta$

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

(b) The class MON_d^n *can be exactly learned with confidence* $1 - \delta$

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

5.5 The Noisy Case: Non-uniform Attribute Distributions

In this section we 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 $d_1, \dots, d_n$. 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 3.2.1). 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 presented in Section 2.2. In this setting, we work with the D-biased inner product $\langle f, g \rangle_D = \mathbb{E}_{x \sim D}[f(x)g(x)]$, the orthonormal basis $\left(\chi_I^D \mid I \subseteq [n]\right)$, and the D-biased Fourier coefficients $\hat{f}(I) = \left\langle f, \chi_I^D \right\rangle_D$, using the same notation as in the uniform case. The fact that Lemma 2.3.4 has been formulated in this D-biased setting paves 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,

$$\chi_I^D(x \oplus \xi) \neq \chi_I^D(x) \cdot \chi_I^D(\xi) .$$

Hence, we cannot simply approximate $\mathbb{E}_{x\sim D,\xi\sim P,\zeta\sim\eta}\left[\chi_I^D(x\oplus\xi)\cdot f(x)\cdot\zeta\right]$ and proceed as in the uniform case. On the other hand, using $\chi_I^{U_n}$, we obtain

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

but $\left\langle f,\chi_I^{U_n}\right\rangle_D$ does not properly work together with the definition of biased Fourier coefficients. The way out is provided by a 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 5.5.1. Its proof relies on explicit calculations of the *biased* Fourier coefficients of the *unbiased* parity functions.

Lemma 5.5.1. *Let $f:\{0,1\}^n\to\mathbb{R}$ and $I\subseteq[n]$. Then*

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

Before we prove Lemma 5.5.1, we calculate $\left\langle\chi_I^{U_n},\chi_J^D\right\rangle_D$. 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 $\left(\chi_I^{U_n}\mid I\subseteq[n]\right)$ to coordinates with respect to the D-biased basis $\left(\chi_I^D\mid I\subseteq[n]\right)$.

Lemma 5.5.2. *Let $J\subseteq I\subseteq[n]$. Then*

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

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

$$\begin{aligned}\left\langle\chi_I^{U_n},\chi_J^D\right\rangle_D &= \sum_{x\in\{0,1\}^n} D(x)\cdot\chi_I^{U_n}(x)\cdot\chi_J^D(x)\\ &= \sum_{x\in\{0,1\}^n}\prod_{i\in[n]}\left(d_i^{x_i}\cdot(1-d_i)^{1-x_i}\right)\cdot\prod_{i\in J:x_i=0}\tfrac{d_i}{\sigma_i}\cdot\prod_{i\in J:x_i=1}\tfrac{1-d_i}{\sigma_i}\cdot\prod_{i\in I\setminus J}(-1)^{x_i}\\ &= \sum_{x\in\{0,1\}^n}\prod_{i\in[n]\setminus J}\left(d_i^{x_i}\cdot(1-d_i)^{1-x_i}\right)\cdot\prod_{i\in J}\sigma_i\cdot\prod_{i\in I\setminus J}(-1)^{x_i}\end{aligned}$$

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

where, analogously to p_I, we define $d_I = \Pr_{x\sim D}\left[\chi_I^{U_n}=-1\right]$ for $I\subseteq[n]$. By Lemma 3.2.5 (applied to D instead of P), $1-2d_I=\prod_{i\in I}(1-2d_i)$. □

Proof of Lemma 5.5.1. We first show that $\langle\chi_I^{U_n},\chi_J^D\rangle_D=0$ for all $J\not\subseteq I$:

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

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

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

Hence,

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

The claim now follows from Lemma 5.5.2. □

5.6 Learning Relevant Attributes from Non-uniformly Distributed Examples

The threshold to recognize nonzero Fourier coefficients is given by the least absolute value of the considered nonzero coefficients. Thus, we define the *Fourier*

Algorithm 5.4 Noisy-Product-Fourier$_d$

```
input $S = ((x_1^k, \dots, x_n^k), y^k)_{k \in [m]}, D, P, \eta, \rho$
$R \leftarrow \emptyset$
$\phi_0 \leftarrow \frac{1}{(1-2\eta)\cdot m} \sum_{k=1}^{m} y^k$
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 $|\beta| \geq \rho/2$
        then $R \leftarrow R \cup \{x_i\}$
output Noisy-Product-Fourier$_d(S, D, P, \eta, \rho) = R$
```

threshold $\mathrm{thr}_D(f)$ of f with respect to D by

$$\mathrm{thr}_D(f) = \min\left\{ \left|\hat{f}(i)\right| \;\middle|\; x_i \in \mathrm{rel}(f) \right\} . \tag{5.2}$$

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, $\mathrm{thr}_D(f) \geq 2^{-|\mathrm{rel}(f)|}$ by Lemma 2.3.6.

For the next theorem, we stick to the notation fixed in Section 5.2, except that S is now assumed to be a *D-distributed* (P, η)-noisy sample of size m.

Theorem 5.6.1. *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 5.3.1. By Lemma 5.5.1,

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

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 3.3.2 that with probability at least $\delta/(2n)$,

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

provided that

$$m \geq 2 \cdot \ln(4n/\delta) \cdot \frac{4(1-2d_i)^2}{(1-2\eta)^2 \cdot \sigma_i^2 \cdot \rho^2} \; . \tag{5.3}$$

Moreover, we have

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

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

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

provided that

$$m \geq 2 \cdot \ln(4n/\delta) \cdot \frac{4}{(1-2p_i)^2 \cdot (1-2\eta)^2 \cdot \sigma_i^2 \cdot \rho^2} \; . \tag{5.4}$$

The number of examples in the claim dominates both numbers given in (5.3) and (5.4). Thus, with probability at least $1 - \delta/n$,

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

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

For monotone concepts, we obtain

Lemma 5.6.2. *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

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

where for $J \subseteq [n]$ 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 (see Definition 2.3.2). If f is monotone, then $f_{x_i=0} \geq f_{x_i=1}$ or $f_{x_i=0} \leq f_{x_i=1}$. 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,

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

We conclude the proof by showing $\sigma_i \geq \min\{d_i, 1-d_i\}$. If $d_i \leq 1/2$, then $\sigma_i = \sqrt{d_i \cdot (1-d_i)} \geq d_i = \min\{d_i, 1-d_i\}$. 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 monotone 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 5.6.3. *Let $f : \{0,1\}^n \to \{-1,+1\}$ be a parity function, i.e., $f = \chi_I$ for some $I \subseteq [n]$. Then*

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

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 3.2.6 and Lemma 3.2.5), then

$$\mathrm{thr}_D(f) \geq 2 \cdot \gamma_c \cdot \theta^{d-1} . \tag{5.5}$$

Proof. Let $i \in \mathrm{rel}(f) = I$. By Lemma 5.5.2,

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

which proves the equation in the claim. To see the inequality (5.5), 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 5.6.4. *(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 5.6.1 and Lemma 5.6.2; part (b) follows from Theorem 5.6.1 and Lemma 5.6.3. Note that a parity function f is uniquely determined by $\mathrm{rel}(f)$. □

5.7 Constructing a Hypothesis from Non-uniformly Distributed Examples

Next we describe how to construct a hypothesis for general concepts. We use Lemma 5.5.1 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

$$\beta_I = \Big((1-2p_I)(1-2\eta)\prod_{i\in I} 2\sigma_i\Big)^{-1} \cdot \frac{1}{m} \cdot \sum_{k=1}^{m} y^k \chi_I(x^k) - \sum_{J \subsetneq I} \prod_{i \in I\setminus J} \frac{1-2d_i}{2\sigma_i} \beta_J \;. \quad (5.6)$$

Finally, we 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 $\hat{f}(J)$ for $J \subseteq I$ are required. This feedback effect leads to a necessary sample size of $2^{\omega(|\operatorname{rel}(f)|)}$. In case that $|1 - 2d_i| \leq \sigma_i$, the following theorem provides upper bounds on the sample and time complexity for learning monotone juntas from product distributed examples in the presence of γ_a-bounded attribute noise and classification noise (with $\eta \neq 1/2$). Note that $|1 - 2d_i| \leq \sigma_i$ if and only if $|1 - 2d_i| \leq 1/\sqrt{5}$, i.e., $d_i \in [0.2764, 0.7236]$.

Theorem 5.7.1. *Let* $f : \{0,1\}^n \to \{-1,+1\}$. *Let* $|1 - 2d_i| \leq 1/\sqrt{5}$ *for all* $x_i \in \operatorname{rel}(f)$ *such that in addition,* $\rho = \operatorname{thr}_D(f) > 0$. *Then* f *can be exactly recovered with confidence* $1 - \delta$ *from* D*-distributed* (P, η)*-noisy samples of size*

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

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

Before we prove Theorem 5.7.1, we show that a suitable hypothesis can be build, provided that the set of relevant variables is already known:

Lemma 5.7.2. *Let* $|1-2d_i| \leq 1/\sqrt{5}$ *for all* $x_i \in \operatorname{rel}(f)$. *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 (5.6). 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 prove by induction on $|I|$ that $|\beta_I - \hat{f}(I)| \leq \epsilon$ with probability at least $1 - \delta$, provided that

$$m \geq 8 \cdot 2^{|I|^2} \cdot \ln\left(2^{|I|+1}/\delta\right) \cdot \gamma_a^{-2|I|} \cdot \gamma_b^{-2} \cdot \epsilon^{-2} . \tag{5.7}$$

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 3.1.2), with probability at least $1 - \delta$,

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

which is clearly dominated by (5.7).

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

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

and

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

The remainder of the proof is a bit technical, so we provide a brief overview first: we show that with probability at least $1-\delta\cdot 2^{-|I|}$, (5.8) holds, and that with probability at least $1-\delta\cdot(1-2^{-|I|})$, (5.9) holds for all $J \subsetneq I$. Putting these things together, we will obtain that with probability at least $1-\delta$, $|\beta_I - \hat{f}(I)| \leq \epsilon$.

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

$$\Big|\psi_I - \big(\prod_{i\in I}2\sigma_i\big)^{-1}\cdot\big\langle f,\chi_I^{U_n}\big\rangle_D\Big| \leq \epsilon/2\ , \tag{5.8}$$

provided that

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

Again, this is dominated by (5.7) since $|1-2d_i| \leq \sigma_i$ implies $\sigma_i \geq 1/\sqrt{5}$ and thus $\big(\prod_{i\in I}2\sigma_i\big)^{-1} \leq (\sqrt{5}/2)^{|I|} \leq 2^{|I|^2}$.

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

$$|\beta_J - \hat{f}(J)| \leq \epsilon\cdot 2^{-|J|-1}\ , \tag{5.9}$$

provided that

$$m \geq 8\cdot 2^{|J|^2}\cdot\ln\big(2^{|I|+1}/\delta\big)\cdot\gamma_a^{-2|J|}\cdot\gamma_b^{-2}\cdot\big(\epsilon\cdot 2^{-|J|}\big)^{-2}\ .$$

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

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

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

$$\begin{aligned} m &\geq 8 \cdot 2^{(|I|-1)^2} \cdot \ln\left(2^{|I|+1}/\delta\right) \cdot \gamma_a^{-2(|I|-1)} \cdot \gamma_b^{-2} \cdot \epsilon^{-2} \cdot 2^{2(|I|-1)} \\ &= 8 \cdot 2^{|I|^2-2|I|+1+2|I|-2} \cdot \ln\left(2^{|I|+1}/\delta\right) \cdot \gamma_a^{-2(|I|-1)} \cdot \gamma_b^{-2} \cdot \epsilon^{-2} \\ &= 4 \cdot 2^{|I|^2} \cdot \ln\left(2^{|I|+1}/\delta\right) \cdot \gamma_a^{-2(|I|-1)} \cdot \gamma_b^{-2} \cdot \epsilon^{-2} \,. \end{aligned}$$

The latter sample bound is again dominated by (5.7). Finally,

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

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 $|\beta_I - \hat{f}(I)| \leq \sqrt{2^{-d} \cdot \epsilon}$ for all $I \subseteq R$. Similarly to the end of the proof of Theorem 3.6.3, the standard LMN analysis (see Linial et al. [LMN93]) yields

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

Let $\epsilon = 2^{-2d}$. Then, with probability at least $1 - \delta$,

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

(note that $\min\{d_i, 1 - d_i\} > 1/4$). This implies $h = f$. Thus, we can request

$$\begin{aligned} m &\geq \operatorname{poly}\Big(2^{d^2}, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1}, 2^{3d/2}\Big) \\ &= \operatorname{poly}\Big(2^{d^2}, \log(1/\delta), \gamma_a^{-d}, \gamma_b^{-1}\Big) \end{aligned}$$

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 5.7.1:

Proof of Theorem 5.7.1. By Theorem 5.6.1, 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 $m \geq \operatorname{poly}(\log n, \log(1/\delta), \gamma_a^{-1}, \gamma_b^{-1}, \gamma_c^{-1}, \rho^{-1})$. By Lemma 5.7.2, f can be exactly recovered from

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

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

Corollary 5.7.3. *If for all $i \in [n]$, $|1 - 2d_i| \leq 1/\sqrt{5}$, then 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*

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

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

The restriction $d_i \in [0.2764, 0.7236]$ may seem a bit unnatural. However, if we allow $|1 - 2d_i|/\sigma_i$ to become arbitrarily large, then for all $J \subsetneq I$, $\hat{f}(J)$ has to be approximated too accurately in order to obtain a good estimate for $\hat{f}(I)$, thus forcing an unreasonably large sample size. One possible way out is to consider the quotient $|1 - 2d_i|/\sigma_i$ as an additional parameter. As the statements of the preceding results are already considerably technical, we forbear from introducing yet another parameter in this chapter.

CHAPTER 6

Learning Parity Juntas from Noisy Data

In the previous two chapters, we have analyzed two groups of algorithms that are based on quite different ideas. Nevertheless, it turned out that the frontier of what can be learned efficiently by these algorithms (or by their extensions) is the same for both paradigms. In particular, both approaches fail to learn the class of parities of up to d variables from uniformly distributed samples unless they consider all sets $|E_I|$ (respectively, Fourier coefficients $\hat{f}(I)$) for $|I| \leq d$. Noticeably, for non-degenerate θ-bounded product attribute distributions (with $\theta > 0$), we could show that PAR_d^n *is* efficiently learnable, see Section 5.5. A similar result is exploited in Feldman's algorithm for (randomized) learning of parities from poly$(\log n)$ unadaptively chosen membership queries [Fel05].

In this chapter, we discuss whether alternative approaches can learn parity juntas from uniformly distributed samples in the presence of noise. The uniform distribution seems to be the hardest input distribution for this problem.

Learning parity juntas under noise splits into two subproblems that already appear to be hard to solve: learning parity juntas from poly$(\log n)$ (even noise-free) examples and learning parities in the presence of noise. Concerning the latter problem, we will see that we can restrict ourselves to studying classification noise only.

While in general, algorithms for learning in the presence of pure classification noise cannot be used to learn in the presence of attribute noise, we present such a reduction for parity juntas. Angluin and Laird [AL88] have proposed the method of minimizing the number of disagreements between the hypothesis and the sample to solve the learning problem under pure classification noise. Thus, we can use this method as a third approach to learn parity juntas in the presence of attribute and classification noise (next to the greedy and the Fourier approach). The fastest known algorithm that implements the disagreement minimization method needs as many steps as d-Greedy or d-Fourier. While the disagreement minimization is applicable in the distribution-free scenario, the attribute distribution has to be known in order to apply the Fourier algorithm. Concerning the greedy algorithm, the attribute distribution does not matter for the purpose of applying it. However, it is not clear for which non-uniform distributions and concepts it succeeds in finding the relevant attributes.

In Section 6.1, we review the state of the art concerning the learnability of parity functions under the constraint of few relevant variables or the presence of noise. The reduction of attribute noise to pure classification noise and the method of minimzing the number of disagreements between the hypothesis and the input data are presented in Section 6.2. In Section 6.3, we study the noise stability of and the noisy distance between parity concepts. The computation of the noisy distance is used to show a lower bound for learning subclasses of parities in Section 6.4. In that chapter, we also present a lower bound that shows that for some γ_a-bounded product distribution P, $\Omega(\gamma_a^{-d})$ examples are needed to learn from uniformly distributed $(P, 0)$-noisy samples.

In this chapter, the parities map to $\{-1, +1\}$, i.e., we deal with the functions χ_I, $I \subseteq [n]$.

6.1 State of the Art

The standard PAC learning algorithm for parities draws a sample

$$S = (x^k, y^k)_{k \in [m]} \in (\{0,1\}^n \times \{-1,+1\})^m$$

of size $m = \Theta(n)$. Then it solves the corresponding system of linear equations

$$Xv = y \tag{6.1}$$

over the two-element field $\mathrm{GF}(2) = \{0, 1\}$, where

$$X = (x_i^k)_{k \in [m], i \in [n]}$$

denotes the $(m \times n)$-matrix of attribute values, i.e., the left part of the matrix representation (3.1) of a sample. There is a one-to-one correspondence between solutions $v \in \{0,1\}^n$ for (6.1) and parity functions χ_I that are consistent with S: $Xv = y$ if and only if χ_I with $I = \{i \in [n] \mid v_i = 1\}$ is consistent with S. With high probability, a consistent parity is ϵ-close to the target parity by Blumer et al. [BEHW87], see also Lemma 3.1.8. Thus, the class PAR^n is learnable distribution-freely from $\Theta(n)$ examples. This approach has been proposed by Helmbold et al. [HSW92].

From a lower bound due to Blumer et al. [BEHW89], it follows that $\Omega(n)$ examples are needed to PAC learn PAR^n (since the VC-dimension of PAR^n is equal to n). As a consequence, the problem of learning arbitrary parities from noise-free examples is completely settled: on the one hand, $\Omega(n)$ examples are needed to learn for information-theoretic reasons. On the other hand, $O(n)$ examples suffice to learn in polynomial time, even in the original distribution-free PAC learning model of Valiant [Val84].

The questions now are:

1. *Can we obtain such tight results also for parities that depend on at most d out of n variables?*
2. *What happens if noise is introduced?*

In the following, we present known partial answers to these questions.

6.1.1 Learning Parity Juntas from Noise-free Data

Let $f = \chi_I \in \mathrm{PAR}_d^n$, i.e., $I \subseteq [n]$ with $|I| \leq d$. If the number of examples is allowed to be $\Theta(n)$, then the algorithm for general parities will do the job. This algorithm is fast, but not attribute-efficient. The size of PAR_d^n is $V(n, d)$ since PAR_d^n consists exactly of the parity functions χ_I with $I \subseteq [n]$ and $|I| \leq d$. By Lemma 3.1.8, any hypothesis $h \in \mathrm{PAR}_d^n$ that is consistent with a sample S for f of size

$$\frac{1}{\epsilon} \ln \frac{|\,\mathrm{PAR}_d^n\,|}{\delta} \leq \frac{d}{\epsilon} \ln \frac{n}{\delta}$$

is ϵ-close to f. Thus, all we have to do for attribute-efficient distribution-free learning is to produce a parity function that is consistent with S. This is equivalent to finding a solution v of (6.1) of Hamming weight at most d. In general, this is the Maximum-Likelihood Decoding problem, which has been shown to be NP-complete by Berlekamp, McEliece, and van Tilborg [BMvT78]. However, in the learning problem, we do not have to cope with worst-case instances, but with random instances. This raises the question for average-case hardness results (see Levin [Lev86] for an introduction to average-case completeness). As far as we know, no such results with implications for the complexity of learning parity juntas are known. In another direction, fixed-parameter tractability results (see Downey and Fellows [DF99]) may yield efficient learning algorithms, even if we assume that $\mathsf{P} \neq \mathsf{NP}$. However, Downey, Fellows, Vardy, and Whittle [DFVW99] have shown that Maximum-Likelihood Decoding is $\mathsf{W}[1]$-hard. This means that it is unlikely that there exists a deterministic algorithm that finds a solution of (6.1) of Hamming weight at most d with worst-case running time $T(d) \cdot n^c$, where c is a constant independent of d and T is some arbitrary function.

If $\mathsf{P} = \mathsf{NP}$, then Maximum-Likelihood Decoding can be solved in polynomial time, and thus, PAR_d^n can be learned from $O(d \log n)$ examples in polynomial time, even in the distribution-free PAC learning model. The weaker assumption that the W-hierarchy collapses with $\mathsf{FPT} = \mathsf{W}[1]$ yields a learning algorithm for learning PAR_d^n from $O(d \log n)$ examples in time $T(d) \cdot n^c$. By contraposition, showing that PAR_d^n cannot be learned efficiently from $O(d \log n)$ examples would imply strong complexity theoretic results such as $\mathsf{P} \neq \mathsf{NP}$ or $\mathsf{FPT} \neq \mathsf{W}[1]$.

Blum et al. [BKW03] have provided a hardness result for a restricted learning paradigm. They have shown that any statistical query algorithm must make $\Omega(n^{d/3})$ queries to learn d-parities. The statistical queries model has been introduced by Kearns [Kea98]. Blum et al.'s lower bound transfers to $\Omega(n^{d/3})$ examples needed for any learning algorithm that is exclusively based on the evaluation of expected values. In particular, the attempt to learn a target parity χ_I by computing $\hat{g}(I) = \langle g, \chi_I \rangle$ for suitable functions g is essentially a statistical query algorithm and hence requires a (too) large number of random examples.

On the positive side, Uehara, Tsuchida, and Wegener [UTW97] have shown how to learn d-parities from $d \log(n/d) + O(d)$ membership queries. This meets the information theoretic lower bound of Turán [Tur93] for this model. The technique of Uehara et al. uses binary search based on so-called splitters. However, it seems crucial that they can query very specific points. Feldman [Fel05] showed how to learn PAR_d^n from $O(d \log n)$ unadaptively chosen membership queries.

Concerning the learnability from random examples, PAR_d^n is efficiently learnable from (strictly) non-uniformly distributed examples, as we have shown in Section 5.5. The lowest known sample bound that allows polynomial-time learning of PAR_d^n from arbitrarily distributed examples is due to Klivans and Servedio [KS06]. They have shown that for $d \in o(\log n)$, PAR_d^n is PAC learnable from a sample of sublinear size in polynomial time. Their algorithm simply ignores a random set of attributes and solves a system of linear equations restricted to the remaining attributes. If there is no solution, the process is repeated. With high probability, all ignored attributes are indeed irrelevant (after a small number of repetitions). The second trick of Klivans and Servedio is to enlarge the hypothesis space from parities of d variables to parities of up to $n^{1-1/d}$ variables. This still requires only sublinearly many examples, but simultaneously permits to output an arbitrary solution of the equation system provided that $n - n^{1-1/d}$ variables have been ignored. On the other hand, the fastest known attribute-efficient algorithm for learning PAR_d^n from arbitrarily distributed examples is due to Spielman (as reported by Klivans and Servedio [KS06]). It has a running time of $n^{d/2} \cdot \text{poly}(n, d)$. For the setting of uniformly distributed examples, no better bounds are known.

6.1.2 Learning Parity Functions from Noisy Data

In the presence of classification noise with rate $\eta < 1/2$, the method of minimizing disagreements proposed by Angluin and Laird [AL88] needs only

$$O(\log(|\mathcal{C}|/\delta)(1-2\eta)^{-2}\epsilon^{-2})$$

examples to learn the class $\mathcal{C}$ with confidence $1-\delta$ and accuracy $1-\epsilon$. However, the task of computing such a minimizing hypothesis is in general computationally hard: it is equivalent to the NP-hard Nearest Codeword problem [ABSS97], which is closely related to Maximum-Likelihood Decoding mentioned above. As we have argued already, the caveat is that we do not really deal with worst-case instances, but with randomly drawn instances with restrictions on the number of relevant attributes. However, there do not seem to be any average-case or fixed-parameter results for this problem yet. Similarly as for learning d-juntas, one can reason that $\mathsf{P} = \mathsf{NP}$ would imply that PAR^n is efficiently learnable in the presence of classification noise. By contraposition, lower sample bounds of size $\omega(d \log n)$ for learning parities from noisy examples would imply strong complexity theoretic results.

Blum, Kalai, and Wasserman [BKW03] showed how to learn parities that depend only on variables among the first k ones from $2^{O(k/\log k)}$ queries with arbitrary classification noise ($\eta < 1/2$). For $k = \log n$, this is an example of a class that is learnable from random noisy examples but not learnable from statistical queries. Their technique is roughly described as follows: from a sufficiently large sample, additional examples of the form $(e_i, \chi_I(e_i) \cdot \zeta)$ are generated as linear combinations of other examples (since we use values $-1, +1$ for classifications, these have to be multiplied). The probability that $\zeta = -1$ is slightly smaller than $1/2$, and thus it can eventually be inferred whether $i \in I$ or not (with high probability). However, it seems that this technique will not lead to improved estimates for d-parities. It seems difficult to design a more efficient algorithm for d-parities based on solving systems of linear equations.

For the class PAR^n, as we will show in Section 6.2, it is possible to reduce product attribute noise to well-behaved classification noise for the case of uniformly distributed attributes and product attribute distributions with rates that are smaller than one half.

To the best of our knowledge, there has been very little research linking parity juntas and learning from noisy data yet. The only result we know of is due to Feldman, Gopalan, Khot, and Ponnuswami [FGKP06]. They have shown that one can reduce the problem of learning d-juntas from uniformly distributed noise-free examples to the problem of learning d-parities from uniformly distributed noisy examples.

6.2 The Disagreement Minimization Method

In this section, let $f = \chi_I : \{0,1\}^n \to \{-1,+1\}$ for some $I \subseteq [n]$, i.e.,

$$f(x) = (-1)^{\sum_{i \in I} x_i}$$

for all $x \in \{0,1\}^n$. Moreover, fix an attribute noise distribution P and a classification noise rate $\eta < 1/2$.

A general method to learn in the presence of pure classification noise has been proposed by Angluin and Laird [AL88], generalizing a result of Blumer et al. [BEHW87]. Angluin and Laird proved that a concept class $\mathcal{C}$ can be learned distribution-freely from $(-,\eta)$-noisy samples with confidence $1-\delta$ and accuracy $1-\epsilon$ from a sample of size

$$m \geq \frac{2}{\epsilon^2(1-2\eta_b)^2} \ln\left(\frac{2|\mathcal{C}|}{\delta}\right) ,$$

where $\eta_b \in [0,1/2)$ is an upper bound on the classification noise rate η. Specifically, any hypothesis minimizing the number of disagreements between the given examples and h's prediction satisfies $\Pr_{x\sim D}[f(x) \neq h(x)] < \epsilon$ with probability at least $1-\delta$.

In the presence of attribute noise, this method fails in general. This is because it can happen that for an example $(x \oplus \xi, f(x))$, one cannot decide whether $\Pr[f(x \oplus \xi) = f(x)]$ is smaller or larger than $1/2$ (or even equal to $1/2$) unless one already knows the target concept f. For the class PAR_d^n, however, it *is* possible to reduce attribute noise to pure classification noise, as we describe in the following.

Lemma 6.2.1. *Let $x \sim U_n$, $\xi \sim P$, and $\zeta \sim \eta$. Then $(x \oplus \xi, f(x) \cdot \zeta)$ is equally distributed as $(x, f(x) \cdot f(\xi) \cdot \zeta)$. Viewing the latter pair as an example with pure classification noise, the classification noise rate is*

$$\eta_I = \tfrac{1}{2}(1 - (1-2p_I)(1-2\eta)) . \tag{6.2}$$

Thus, $\eta_I \neq 1/2$ if and only if $p_I \neq 1/2$ and $\eta \neq 1/2$.

Proof. By Lemma 3.2.4, if $x \sim U_n$, $\xi \sim P$, and $\zeta \sim \eta$ are independent random variables, then $(x \oplus \xi, f(x) \cdot \zeta)$ and $(x, f(x \oplus \xi) \cdot \zeta)$ are identically distributed.

Since for $f = \chi_I$, $f(x \oplus \xi) = f(x) \cdot f(\xi)$, we may consider $(x, f(x \oplus \xi) \cdot \zeta) = (x, f(x) \cdot f(\xi) \cdot \zeta)$ as an example with pure classification noise. The classification noise bit is then $f(\xi) \cdot \zeta$, the distribution of which is determined by

$$\begin{aligned} \Pr_{\xi \sim P, \zeta \sim \eta}[f(\xi) \cdot \zeta = -1] &= p_I(1-\eta) + (1-p_I)\eta \\ &= \tfrac{1}{2}\left(1 - (1-2p_I)(1-2\eta)\right) . \end{aligned}$$

□

The next theorem states under which conditions the disagreement minimization method is successful for concept classes $\mathcal{C} \subseteq \mathrm{PAR}^n$. Recall that $\lambda_I = 1 - 2p_I$ is the eigenvalue of the noise operator T_P associated with the eigenfunction χ_I.

Theorem 6.2.2. *Let $\mathcal{C} \subseteq \mathrm{PAR}^n$ and $\delta, \epsilon > 0$. Let P be an attribute noise distribution such that $p_I < 1/2$ for all $I \subseteq [n]$ with $\chi_I \in \mathcal{C}$. Let*

$$\lambda = \min\{|\lambda_I| \mid I \subseteq [n] \wedge \chi_I \in \mathcal{C}\} . \tag{6.3}$$

The disagreement minimization method distribution-freely learns the class $\mathcal{C}$ with confidence $1-\delta$ and accuracy $1-\epsilon$ from

$$m \geq \frac{2}{\epsilon^2 \lambda^2 (1-2\eta)^2} \ln\left(\frac{2|\mathcal{C}|}{\delta}\right)$$

(P, η)-noisy examples.

Proof. Let

$$\eta_b = \max\{\eta_I \mid I \subseteq [n] \wedge \chi_I \in \mathcal{C}\}$$

with η_I as defined in (6.2). Then

$$1 - 2\eta_b \geq \min\{(1-2p_I) \cdot (1-2\eta) \mid I \subseteq [n] \wedge \chi_I \in \mathcal{C}\} = \lambda \cdot (1-2\eta) .$$

The claim now follows from Angluin and Laird's result described in the beginning of this section. □

The advantage of the disagreement minimization method is that it requires no a priori knowledge about the attribute distribution. For uniformly distributed examples, we obtain

Corollary 6.2.3. *Let P be a product distribution such that there exists a constant $\gamma_a > 0$ with $p_i < (1-\gamma_a)/2$ for all $i \in [n]$. Let $\delta, \epsilon > 0$. Then the disagreement minimization method exactly learns the class PAR_d^n with confidence $1-\delta$ from*

$$m \geq d \cdot 2^{2d+3} \cdot \ln(2n/\delta) \cdot \gamma_a^{-d} \cdot (1-2\eta)^{-2}$$

uniformly distributed (P, η)-noisy examples.

Proof. As we have mentioned before, exact learning of d-juntas from uniformly distributed examples is equivalent to learning with accuracy $\epsilon = 2^{-d-1}$, see Lemma 2.3.7. Furthermore, $|\,\mathrm{PAR}_d^n\,| \leq n^d$. Finally, the restrictions on P guarantee that $\lambda_I \geq \gamma_a^{|I|}$ for all $I \subseteq [n]$. Specifically,

$$\min\{\lambda_I \mid I \subseteq [n] \wedge |I| \leq d\} \geq \gamma_a^d \ .$$

The claim follows from Theorem 6.2.2. □

As mentioned before, there are no algorithms available to solve disagreement minimization efficiently. The fastest method seems to be a brute-force search through all n-ary parities of up to d variables to find some disagreement minimizing hypothesis. This takes $n^d \cdot \mathrm{poly}(m, n)$ steps in the worst case. Thus, the performance of disagreement minimization is identical to the performance of d-Greedy or d-Fourier for uniformly distributed (P, η)-noisy examples. However, disagreement minimization works distribution-freely, whereas the Fourier method only works for a known attribute distribution. On the other hand, to successfully apply disagreement minimization, we have to require that $p_I < 1/2$ for all $\chi_I \in \mathcal{C}$, whereas such a restriction is not necessary for the other methods.

6.3 Noise Stability versus Noisy Distance of Parity Functions

We first compute the noise stability of a class $\mathcal{C} \subseteq \mathrm{PAR}^n$. Recall that the noise stability of $\mathcal{C}$ with respect to the attribute noise distribution P is defined as

$$\Gamma_P(\mathcal{C}) = \min\{\Gamma_P(f) \mid f \in \mathcal{C}\} \ ,$$

where

$$\Gamma_P(f) = \min\left\{ \left|1 - 2 \Pr_{\xi \sim P}[f(x \oplus \xi) \neq f(x)]\right| \ \middle| \ x \in \{0,1\}^n \right\} \ .$$

Lemma 6.3.1. *Let $\mathcal{C} \subseteq \mathrm{PAR}^n$. Then $\Gamma_P(\mathcal{C}) = \lambda$ with λ as defined in (6.3).*

Proof. By (3.12), $\Gamma_P(f) = \min\{|T_P(f)(x)| \mid x \in \{0,1\}^n$. By Lemma 3.4.2 (a), $T_P(\chi_I) = \lambda_I \cdot \chi_I$. Consequently, $|T_P(\chi_I)(x)| = |\lambda_I|$ for all $x \in \{0,1\}^n$ and thus, $\Gamma_P(f) = \lambda$. □

In other words: the noise stability $\Gamma_P(\mathcal{C})$ is equal to the smallest eigenvalue of the noise operator T_P (in absolute value). The lemma shows that if $\lambda_I \geq 0$ for all $I \subseteq [n]$ with $\chi_I \in \mathcal{C}$, then the performance of the disagreement minimization method depends on the noise stability. For $\mathcal{C} = \mathrm{PAR}_d^n$, the same is true for the Fourier algorithm since we have to be able to divide by $1 - 2p_I$ for all $I \subseteq [n]$ with $|I| \le d$.

Is there a noise distribution such that PAR_d^n is in principle learnable but not via disagreement minimization or the Fourier algorithm? We answer this question affirmatively by comparing the noise stability $\Gamma_P(\mathrm{PAR}_d^n)$ and the noisy distance $\Delta_P^\epsilon(\mathrm{PAR}_d^n)$. As we have proved in Section 3.7, learning in the presence of noise is possible if and only if $\Delta_P^\epsilon(\mathcal{C}) > 0$. Let us first show that the exact choice of the accuracy parameter ϵ plays no role:

Lemma 6.3.2. *Let $\mathcal{C} \subseteq \mathrm{PAR}^n$. Then, for all $\epsilon \in [0, 1/2)$, $\Delta_P^\epsilon(\mathcal{C}) = \Delta_P^0(\mathcal{C})$.*

Proof. It is clear from the definition that Δ_P^ϵ decreases as ϵ decreases. Therefore, it remains to show that $\Delta_P^\epsilon(\mathcal{C}) \le \Delta_P^0(\mathcal{C})$. Let $I, J \subseteq [n]$ with $\chi_I, \chi_J \in \mathcal{C}$ and $\Delta_P(\chi_I, \chi_J) > 0$. Then $I \neq J$, and χ_I and χ_J differ in exactly half of all inputs $x \in \{0,1\}^n$. This is because $\chi_I(x) \neq \chi_J(x)$ if and only if $\chi_{I \triangle J}(x) = -1$. Unless $I = J$, the latter happens with probability 1/2 for $x \sim U_n$. Since $\epsilon < 1/2$, $\Delta_P^\epsilon(\mathcal{C}) \le \Delta_P(\chi_I, \chi_J)$. Consequently, $\Delta_P^\epsilon(\mathcal{C}) \le \Delta_P^0(\mathcal{C})$. □

Now we compute the noisy distance Δ_P^ϵ between two parity functions χ_I and χ_J.

Lemma 6.3.3. *Let $I, J \subseteq [n]$ with $I \neq J$ and $P : \{0,1\}^n \to [0,1]$ be a probability distribution. Then*

$$\Delta_P(\chi_I, \chi_J) = \tfrac{1}{2} \max\{|\lambda_I|, |\lambda_J|\} .$$

Proof.

$$\begin{aligned}
\Delta_P(\chi_I, \chi_J) &= \tfrac{1}{2} \sum_x \mathbb{E}_{x \sim U_n}[|\mathbb{E}_{\xi \sim P}[\chi_I(x \oplus \xi) - \chi_J(x \oplus \xi)]|] \\
&= 2^{-n-1} \sum_{x \in \{0,1\}^n} \Big| \sum_{\xi \in \{0,1\}^n} P(\xi)(\chi_I(x \oplus \xi) - \chi_J(x \oplus \xi)) \Big| \\
&= 2^{-n-1} \Big(\sum_{x : \chi_I(x) = \chi_J(x)} \Big| \sum_{\xi \in \{0,1\}^n} P(\xi)(\chi_I(\xi) - \chi_J(\xi)) \Big| \\
&\qquad + \sum_{x : \chi_I(x) \neq \chi_J(x)} \Big| \sum_{\xi \in \{0,1\}^n} P(\xi)(\chi_I(\xi) + \chi_J(\xi)) \Big| \Big) .
\end{aligned}$$

Since $I \neq J$, $\chi_I(x) = \chi_J(x)$ for exactly half of the $x \in \{0,1\}^n$. Consequently,

$$\Delta_P(\chi_I, \chi_J) = \tfrac{1}{4}(|\lambda_I - \lambda_J| + |\lambda_I + \lambda_J|) = \tfrac{1}{2} \max\{|\lambda_I|, |\lambda_J|\} .$$

□

Corollary 6.3.4. *Let $\mathcal{C} \subseteq \mathrm{PAR}^n$ with $|\mathcal{C}| \geq 2$ and $\epsilon \in [0, 1/2)$. Then*

$$\Delta_P^\epsilon(\mathcal{C}) = \tfrac{1}{2} \min \left\{ \max\{|\lambda_I|, |\lambda_J|\} \;\middle|\; I, J \subseteq [n], \chi_I, \chi_J \in \mathcal{C}, I \neq J \right\} ,$$

i.e., $\Delta_P^\epsilon(\mathcal{C})$ is equal to half the second smallest $|\lambda_I|$ of the functions χ_I in $\mathcal{C}$.

In particular, $2\Delta_P^\epsilon(\mathrm{PAR}^n)$ is equal to the second smallest eigenvalue of the noise operator T_P (in absolute value). If P is a γ_a-bounded distribution (see Definition 3.2.6), then

$$2\Delta_P^\epsilon(\mathrm{PAR}_d^n) \geq \Gamma_P(\mathrm{PAR}_d^n) \geq \gamma_a^d .$$

In the proof of the following theorem, we construct a simple concept class $\mathcal{C}$ and an attribute noise distribution P for which $\Delta_P^\epsilon(\mathcal{C}) > 0$ but $\Gamma_P(\mathcal{C}) = 0$ (in particular, P cannot be a product distribution). For such a class, the method of minimizing the number of disagreements fails although $\mathcal{C}$ is in principle learnable from uniformly distributed $(P, 0)$-noisy samples by Theorem 3.7.2.

Theorem 6.3.5. *There is a concept class $\mathcal{C} \subseteq \mathrm{PAR}^n$ and an attribute noise distribution P such that $\Delta_P^0(\mathcal{C}) > 0$, but $\Gamma_P(\mathcal{C}) = 0$.*

Proof. Let $P(0^n) = 1/2$ and $P(e_i) = \frac{1}{2n}$ for $i \in [n]$. Let $\mathcal{C} = \{\chi_\emptyset, \chi_{[n]}\}$. Then $\lambda_\emptyset = 1/2$ and $\lambda_{[n]} = 0$. Hence $\Delta_P^0 = 1$ by Corollary 6.3.4 and $\Gamma_P(\mathcal{C}) = 0$ by Lemma 6.3.1. □

In the previous proof, for the target concept $\chi_{[n]}$, the disagreement minimization algorithm will output any of the two admissible concepts with probability 1/2. In contrast, the concept $\chi_\emptyset$ will be correctly identified almost surely. A trivial learning algorithm for $\mathcal{C}$ classifies samples constantly labeled by ones as originating from the concept $\chi_\emptyset$ and all other samples as coming from the concept $\chi_{[n]}$. The probability of misclassification is 2^{-n} in case that the target concept is $\chi_{[n]}$ and 0 otherwise. Thus, we only need $O(\log(1/\delta))$ examples to exactly learn the class $\mathcal{C}$ with confidence $1 - \delta$.

6.4 Lower Bounds via the Noisy Distance

Bshouty et al. [BJT03, Theorem 6] proved that if $P(\xi)$ is superpolynomially small for all $\xi \in \{0, 1\}^n$, then $\Delta_P^\epsilon(\mathrm{PAR}^n)$ is also superpolynomially small, yielding that PAR^n is not learnable from uniformly distributed $(P, 0)$-noisy samples of polynomial size. By Corollary 6.3.4, this means that for any such probability distribution P, there exists a set $I \subseteq [n]$ such that $\Pr_{\xi \sim P}[\chi_I(\xi) = -1]$ is superpolynomially close to 1/2. We now show that the result of Bshouty et al. can be directly inferred from the calculation of $\Delta_P^\epsilon(\mathrm{PAR}^n)$, thus avoiding the detour made in the original proof [BJT03] via the α-attenuated power spectrum s_P (as defined in (3.11), see also [BJT03]). However, our computation also needs some Fourier analysis.

Theorem 6.4.1. *Let $\mathcal{C} \subseteq \mathrm{PAR}^n$ with $|\mathcal{C}| \geq 2$ and $\epsilon \in [0, 1/2)$. Then*

$$\Delta_P^\epsilon(\mathcal{C}) \leq \frac{1}{2} \cdot \left(\frac{2^n}{|\mathcal{C}| - 1} \cdot \max\left\{P(\xi) \mid \xi \in \{0,1\}^n\right\} \right)^{1/2} .$$

Proof. Recall that $\lambda_I = 2^n \hat{P}(I)$ (see (3.7). In the following, all norms are taken with respect to the uniform distribution. By Parseval's equation (2.10), we have

$$2^{-n} \cdot \sum_{\xi \in \{0,1\}^n} P(\xi)^2 = \|P\|_2^2 = \sum_{I \subseteq [n]} \hat{P}(I)^2$$

and

$$\|P\|_2^2 = 2^{-n} \cdot \sum_{\xi \in \{0,1\}^n} P(\xi)^2 \leq 2^{-n} \cdot \max_{\xi \in \{0,1\}^n} P(\xi) \cdot \sum_{\xi \in \{0,1\}^n} P(\xi) = 2^{-n} \cdot \max_{\xi \in \{0,1\}^n} P(\xi) .$$

Let $J \subseteq [n]$ such that $|\hat{P}(J)|$ is the second smallest among all $|\hat{P}(I)|$ with $\chi_I \in \mathcal{C}$. Then

$$\hat{P}(J)^2 \leq \frac{\|P\|_2^2}{|\mathcal{C}| - 1}$$

since otherwise $\|P\|_2^2 \geq 0 + (|\mathcal{C}| - 1) \cdot \hat{P}(J)^2 > \|P\|_2^2$. Now

$$2 \cdot \Delta_P^\epsilon(\mathcal{C}) = |\lambda_J| = 2^n \cdot |\hat{P}(J)| \leq \frac{2^n}{(|\mathcal{C}| - 1)^{1/2}} \cdot \|P\|_2 \leq \left(\frac{2^n}{|\mathcal{C}| - 1} \cdot \max_{\xi \in \{0,1\}^n} P(\xi) \right)^{1/2} .$$

□

Setting $\mathcal{C} = \mathrm{PAR}^n$ in the previous theorem, we obtain

$$\Delta_P^\epsilon(\mathrm{PAR}^n) \leq (1 + o(1)) \cdot \left(\max_{\xi \in \{0,1\}^n} P(\xi) \right)^{1/2}$$

and hence recover the result of Bshouty et al. [BJT03, Theorem 6] that the number of examples needed to learn the class of parity functions from uniformly distributed examples under attribute noise is inversely polynomially related to the highest probability of a noise vector. For instance, if P is a product distribution with rates $p_i \in [\rho, 1 - \rho]$ for some constant $\rho \in (0, 1/2]$ that is independent of n, then

$$\max_{\xi \in \{0,1\}^n} P(\xi) \leq (1 - \rho)^n$$

is exponentially small in n.

We close this chapter by showing that the factor γ_a^{-d} that appears in all of our positive noise-tolerant learning results is in fact necessary:

Theorem 6.4.2. *Let P be a product distribution with rates*

$$p_1 = \cdots = p_n = p \in [0,1] \setminus \{1/2\} \ .$$

Let $\gamma_a = |1-2p|$. Then P is γ_a bounded and $\Omega(\gamma_a^{-d})$ examples are necessary to learn PAR_d^n from uniformly distributed $(P,0)$-noisy examples. The same is true for arbitrary superclasses of PAR_d^n such as $\mathcal{J}_d^n$.

Proof. Clearly, P is γ_a-bounded. Let $I, J \in [n]$ with $|I| = |J| = d$ and $I \neq J$. Then

$$\Delta_P^0(\chi_I, \chi_J) = \max\{|\lambda_I|, |\lambda_J|\} = |1-2p|^d = \gamma_a^d \ .$$

Thus, $\Delta_P^0(\mathrm{PAR}_d^n) = \gamma_a^d$. By Theorem 3.6.1, $\Omega(\gamma_a^{-d})$ examples are needed to learn PAR_d^n from uniformly distributed $(P,0)$-noisy samples. For a superclass $\mathcal{C} \supseteq \mathrm{PAR}_d^n$, we have $\Delta_P^0(\mathcal{C}) \leq \Delta_P^0(\mathrm{PAR}_d^n)$. Hence, also $\mathcal{C}$ requires $\Omega(\gamma_a^{-d})$ examples to be learned successfully. □

CHAPTER 7

Conclusion

We have investigated two approaches to learn the relevant attributes of a target concept in the presence of attribute and classification noise. On the one hand, we have presented a mathematically sound characterization of a simple greedy algorithm that has been successfully used in practice. Our analysis provides an accurate way of determining whether the greedy algorithm can learn the relevant attributes of a given target concept or not, yielding a sharp dichotomous result. In spite of their importance (at least from a theoretical point of view), results of this type seem quite rare in the literature. We have seen that the greedy approach can be extended to cope with the class of τ-Fourier accessible functions. In addition, the greedy algorithm has turned out to be very robust against noise present in the input data.

On the other hand, we have developed Fourier-based algorithms for learning the class of τ-low juntas from few examples in time roughly n^τ. While the design of such an algorithm is straightforward for the noise-free setting, we have seen that it takes some "detours" to extend the ideas to the noisy scenario. Considering τ as a parameter, we have shown that the sequence of classes of τ-low d-juntas, $1 \leq \tau \leq d$, is learnable by the sequence of algorithms τ-FOURIER (respectively, their noise-tolerant counterparts). Already for $\tau = 1$, many important concepts can be addressed. For $\tau = 2$, interesting concepts such as the *not all equal* function are added to the set of feasible concepts. As τ increases, more and more d-juntas can be learned (in time n^τ), reaching the class of all d-juntas for $\tau = d$.

Moreover, we have seen that a generalization to cope with non-uniformly distributed data is possible—albeit with quite a little effort. In particular, we have shown that monotone and parity juntas are efficiently learnable from few noisy examples when the attributes are non-uniformly distributed.

Overall, we could show that the class of all d-juntas is *in principle* learnable under almost arbitrary attribute and classification noise, the running time being roughly n^d. While this was our initial goal for the Fourier-based approach, it seemed not at all clear that the greedy approach would yield algorithms that are applicable to the same classes of concepts. A major research goal for the future is to devise noise-tolerant learning algorithms for arbitrary d-juntas that run in time $n^{c \cdot d}$ for some constant $c < 1$.

In our Fourier-based approach for finding the relevant attributes, we search for nonzero Fourier coefficients up to level τ for some $\tau \in [d]$, thus being able to learn the class of τ-low concepts. At a first glance, one may thus think that the Fourier approach fails for concepts that are τ-accessible but not τ-low. This is true for τ-Fourier, but a slight modification of the algorithm fixes this shortcoming: each time a relevant variable x_i is identified, the algorithm can recurse for the subconcepts $f_{x_i=0}$ and $f_{x_i=1}$. It is not hard to see that this modified Fourier algorithm exactly learns the class of τ-Fourier-accessible concepts. For the noise-free case, this algorithm has (implicitly) been proposed by Mossel et al. [MOS04].

To learn the class PAR_d^n of n-ary parity functions of up to d variables from noisy examples, we have proposed to use the method of minimizing disagreements between the sample and the hypothesis. This method had been introduced by Angluin and Laird [AL88] to cope with pure classification noise. To use it in our noise scenario, we had to reduce attribute noise to pure classification noise. Although the idea of this method is entirely different from the greedy method and the Fourier method, it is not known whether it can be implemented more efficiently since the underlying minimization problem is generally believed to be computationally hard. The best upper bound on the running time that we can provide is roughly n^d.

We have compared the noise stability and the noisy distance of subclasses of parity juntas. It has turned out that all methods proposed may fail to learn parity juntas in spite of principle learnability. In contrast, these methods are successful if the attribute noise distribution is γ_a-bounded. We have generalized a lower bound of Bshouty et al. [BJT03]. Finally, we have shown a lower bound of $\Omega(\gamma_a^{-d})$ on the number of examples to learn PAR_d^n, showing that the corresponding factor in our positive learning results cannot be improved.

Apart from learning juntas, we have proved a generalization of Bshouty et al.'s result for a noise-tolerant LMN-style algorithm [BJT03]. In addition, we have provided a characterization for general learnability of concept classes in the presence of random attribute and classification noise.

7.1 Multi-valued Attributes and Classifications

It is straightforward to generalize the greedy algorithm to settings in which attributes and classifications take more than two values. However, it is not clear how to analyze the algorithm in this case. Instead of looking at properties of Fourier spectra, it seems more reasonable to consider correlations between the target concept and certain functions of the attributes. These notions do not coincide any more for functions $f : \{0, \ldots, r-1\}^n \to \mathbb{R}$ with $r \geq 3$. Thus, also the Fourier approach should be recast as a purely statistical approach. In this way, however, we lose some algebraic structure. As a consequence, it is not clear whether an algorithm with nontrivial running time can be constructed as in the case $r = 2$. The exhaustive search algorithm still runs in roughly n^d steps for arbitrary r.

Alternatively, if the classification is non-Boolean but still has finite range, one can code this range by a set of Boolean classifications. The relevant variables may then be inferred by learning the relevant attributes of each subconcept corresponding to one classification bit.

7.2 Fourier Transform and Group Representations

In the beginning of Section 2.2, we have presented the set of Dirac functions as a natural candidate for an othonormal basis of $\mathbb{R}^{\{0,1\}^n}$ with respect to a natural inner product. We have then argued that the Hadamard basis is more useful for us. The whole Fourier analysis on the hypercube (under uniform distribution) was based on this basis. One may now step back and ask: "Wouldn't it make sense to consider different orthonormal bases? What is so special about the Hadamard basis?" One answer is of course that "it works remarkably well and helps to solve a lot of problems." Nevertheless, we will present another answer from a more structural standpoint: we describe why the Hadamard basis is a canonical choice and how it fits into the very general framework of group representations and character theory. The investigation of different orthonormal bases may nonetheless be a worthy aspect of future research.

In a certain sense, the theory of Fourier analysis on the hypercube unfolds its entire beauty only when considered from a higher standpoint, i.e., if one embeds it into the theory of complex group representations. For our purposes, we are concerned with the Abelian group $G = \mathbb{Z}_2^n$. The *dual group* or *character group* $\hat{G}$ is the set of group homomorphisms from G to $\mathbb{C}^*$, the multiplicative group of nonzero complex numbers. Equipped with pointwise multiplication, the set $\hat{G}$ becomes an Abelian group. Since for finite G, all elements of G

are of finite order, any such homomorphism maps into the group $\mathbb{T}$ of complex numbers with modulus one. Moreover, since all elements of $\mathbb{Z}_2^n$ are of degree 2, these homomorphisms even map to $\{-1, +1\}$, the group of units in $\mathbb{R}$. Let $\rho : G \to \mathbb{C}^*$ be a group homomorphism. An element $\rho(g)$ can be viewed as being a multiplication factor and thus affecting the complex plane by a combination of a dilation and a rotation. If G is finite, then the dilation factor is 1, and $\rho(g)$ becomes an element of $U(1)$, the group of unitary operations on $\mathbb{C}^1$. In general, a *group representation* of a group G is a group homomorphism $\rho : G \to \mathrm{GL}(\mathbb{C}^n)$. This allows one to interpret elements g from G as linear isomorphisms on $\mathbb{C}^n$. The *dimension* of ρ is defined to be n. The representation ρ is called *irreducible* if $\mathbb{C}^n$ has no non-trivial G-invariant subspaces, i.e., whenever we have a subspace $U \subseteq \mathbb{C}^n$ with $\rho(g)(u) \in U$ for all $g \in G$ and all $u \in U$, then $U = \{0\}$ or $U = \mathbb{C}^n$. By Maschke's Theorem, every representation of a finite group G can be decomposed into a direct sum of irreducible representations. The *character* χ_ρ of ρ is defined by $\chi_\rho : G \to \mathbb{C}$,

$$\chi_\rho(g) = \mathrm{tr}(\rho(g)) \text{ for } g \in G \ ,$$

where tr denotes the *trace* of the homomorphism $\rho(g)$, i.e., the sum of the diagonal entries in any matrix representation of $\rho(g)$. Characters are helpful in classifying group representations up to isomorphisms since two representations with the same character are isomorphic. Moreover, if a representation is decomposed into irreducible representations $\rho_1, \ldots, \rho_r$, then the corresponding characters χ_{ρ_i} are orthonormal with respect to the inner product

$$\langle f_1, f_2 \rangle = \frac{1}{|G|} \sum_{g \in G} f_1(g) \overline{f_2(g)}$$

for functions $f_1, f_2 : G \to \mathbb{C}$. Note that this inner product is induced by the uniform distribution μ on G, i.e., $\langle f_1, f_2 \rangle = \int_G f_1 \overline{f_2} d\mu$. In general, one can define a canonical inner product by means of the *(normalized) Haar measure* on locally compact groups. This is the unique measure that satisfies certain properties such as invariance under left-translation by elements from G and regularity.

For one-dimensional representations, $\chi_\rho(g)$ is the only matrix entry of the matrix representation of $\rho(g)$. It can be shown that for Abelian G, every irreducible representation is one-dimensional. Thus, in this case, we can interpret the elements of the character group $\hat{G}$ as *the irreducible characters* of G. In this sense, for the uniform distribution, the Fourier expansion formula (2.8) describes how to decompose an arbitrary complex-valued function on the hypercube into a linear combination of the irreducible characters of $\mathbb{Z}_2^n$. In order to deal with non-uniform distributions, we have to abandon the translation invariance of the measure and adjust the theory accordingly.

For a brief and accessible introduction to group representations, we recommend Artin [Art91, Chapter 9].

7.3 Open Problems

7.3.1 The Greedy Method

The first issue left for future research is a deeper investigation of the greedy algorithm and its variants for non-uniformly distributed attributes.

The second issue is to stick to the scenario of uniformly distributed attributes and investigate further variants of the greedy heuristic: for which functions can greedy algorithms that use a different weighting scheme find the relevant variables? In our case, the weight of variable x_i is equal to the number of edges in the functional relations graph that can be covered by x_i. However, if an edge is labeled by exactly one variable, then this variable has to be selected in order to explain the sample. For this reason, Almuallim and Dietterich [AD94] proposed to assign the weight $\sum_{e \in E_i} \frac{1}{|c(e)|-1}$ to x_i and then find a set cover by selecting variables of maximum weight. Since for $n \gg |\operatorname{rel}(f)|$, each edge is labeled by roughly $n/2$ irrelevant variables, such a weighting is unlikely to help much during the first rounds of the algorithm. Consequently, it is not clear whether the class of functions for which this heuristic succeeds becomes any larger.

7.3.2 The Fourier Method

Concerning the Fourier approach, the canonical open problem is to improve the bounds on sample complexity and running time in case of non-uniformly distributed attributes. Furthermore, one could try to embed $\{0,1\}^n$ into a group different from $\mathbb{Z}_2^n$ and consider the characters associated with the irreducible representations of that group.

Another direction is to impose constraints on the knowledge of the attribute distribution. While distribution-free PAC learning of n-ary d-juntas seems to be hard even in the noise-free case, it may well be that partial knowledge on the distribution suffices in many situations. However, the case of parity juntas, which are learnable under non-uniform distribution but apparently difficult to learn under uniform distribution, seems to draw limits.

7.3.3 Learning Parity Juntas from Noisy Data

The major open question certainly is “Are parity juntas really hard to learn attribute-efficiently under uniform distribution?” If so, this would mean a strong dichotomic result since we have shown that parity juntas are learnable from noisy examples that are generated by product distributions with rates bounded away from $1/2$. For another positive learning result, consider the function $f\colon \{0,1\}^n \to \{0,1\}$ defined by $f(x) = 1$ if $|x| \bmod k \neq 0$ and $f(x) = 0$ otherwise. If $f = 2$, then f is the parity function. If $k \geq 3$, then f is 1-low under the uniform distribution and thus efficiently learnable. More strongly, it may well be that all symmetric concepts are 3-low unless they are parity functions. Approaching from different learning models, learning parity juntas from membership queries is easy, even if the queries have to be chosen non-adaptively in advance [Fel05].

The problem of learning parities from uniformly drawn noisy examples is related to notoriously open problems in the theory of error-correcting codes. Advances in the field of average case complexity and its relationship to worst case complexity may shed more light on these issues.

7.3.4 Beyond Known Methods

The hope is to find an algorithm that significantly outperforms τ-Greedy and τ-Fourier insofar as it runs in polynomial time and is able to learn the relevant attributes of concepts that are not $O(1)$-Fourier-accessible. For noise-free learning of parity juntas (which are not even $(d-1)$-Fourier-accessible), such algorithms are known, at least if one drops the requirement of attribute-efficiency (see Chapter 6). However, if $\tau \in \Theta(n)$, but $\tau < n$, then it is not known whether the class of τ-low juntas is efficiently learnable at all. See Blum [Blu03] for concrete candidates of juntas that seem to be hard to learn already from noise-free data. For the case of noise-affected data, any algorithm with running time $n^{c \cdot d}$ with $c < 1$ would be a substantial progress.

List of Algorithms

List of Definitions

List of Symbols

A_I^{ab}	set of example indices k with $x_I^k = a$ and $y^k = b$, $a, b \in \{0,1\}$; p. 61
$\mathrm{acc}(f)$	set of accessible variables w.r.t. f; p. 29, Definition 2.4.6
$\mathcal{C}$	concept class (of concepts $f : \{0,1\}^n \to \Omega$); p. 15
$\mathbb{C}$	set of complex numbers
$c(e)$	characteristic vector of an edge e of the functional relations graph; p. 56, Equation (4.1)
$\mathrm{Cov}[X, Y]$	covariance of random variables X and Y, p. 21
D	attribute distribution $D : \{0,1\}^n \to [0,1]$; p. 33
d	number of relevant attributes
d_i	rate for x_i if D is a product distribution; p. 37
E	edge set of functional relations graph; p. 56, Definition 4.1.1
E_i	set of edges of functional relations graph that can be covered by attribute x_i; p. 56
E_I	same as E_i for parity variable x_I; p. 59
$E_i^{(s)}$	remaining edges that can be covered by x_i after s rounds of Greedy; p. 64
$E_I^{(s)}$	same as $E_i^{(s)}$ for parity variable x_I; p. 70
$\mathbb{E}[X]$	expectation of random variable X
$\tilde{\mathbb{E}}_S[X]$	empirical expectation of random variable X given S; p. 34, Definition 3.1.3
e	Euler constant $e \approx 2.7183$
e_i	i^{th} unit vector; p. 16
$\exp$	exponential function with base e
f	target concept $f : \{0,1\} \to \Omega$; p. 33
f'	base function of f; p. 23, Definition 2.3.2
$\hat{f}$	Fourier transform of f; p. 19
$\mathcal{F}_D(f)$	Fourier transform of f (w.r.t. distribution D); p. 19

$\tilde{f}_S(I)$	empirical Fourier coefficient of f at I given S; p. 35, Definition 3.1.5
f_a	restriction of f to points with fixed assignment $a \in \{0,1\}^I$; p. 17
$\mathrm{FSG}(f)$	Fourier support graph of f; p. 21, Definition 2.2.1
$\mathrm{GF}(2)$	field of two elements
G_S	functional relations graph of a sample S; p. 56, Definition 4.1.1
h	hypothesis $h : \{0,1\} \to \Omega$; p. 36, Definition 3.2.2
i, j	mostly: variable indices
I, J	subsets of $[n]$
$\mathcal{J}_d^n$	class of n-ary d-juntas; p. 23, Definition 2.3.3
$\mathcal{J}_d^n(\tau)$	class of n-ary τ-low d-juntas; p. 26, Definition 2.4.1
$\mathrm{inacc}(f)$	set of inaccessible variables w.r.t. f; p. 29, Definition 2.4.6
$\mathrm{irrel}(f)$	set of irrelevant variables of f; p. 23, Definition 2.3.1
k, ℓ	mostly: example indices
$\ln$	natural logarithm (base e)
$\log$	binary logarithm (base 2)
m	sample size (number of examples)
MON^n	class of monotone functions $f : \{0,1\}^n \to \Omega$; p. 16
MON_d^n	class of monotone d-juntas $f : \{0,1\}^n \to \Omega$; p. 23
n	number of attributes
$\mathbb{N}$	set of nonnegative integers $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$
P	attribute noise distribution; p. 36
PAR^n	class of parity functions $f : \{0,1\}^n \to \Omega$; p. 16
PAR_d^n	class of parity d-juntas $f : \{0,1\}^n \to \Omega$; p. 23
p_i	short for $p_{\{i\}}$; also: rate for ξ_i if P is a product distribution; p. 37
p_I	p. 38, Equation (3.4)
$\mathcal{P}([n])$	power set of $[n]$, i.e., set of all subsets of $[n]$
$p_\lambda(n, d)$	probability that GREEDY λ-fails for non-Fourier-accessible n-ary d-juntas; p. 67, Corollary 4.3.5
$p^{(\tau)}(n, d)$	probability that τ-GREEDY fails for non-τ-Fourier-accessible n-ary d-juntas; p. 72, Corollary 4.4.5
$\Pr$	probability
$\mathbb{R}$	set of real numbers
$\mathrm{rel}(f)$	set of relevant variables of f; p. 23, Definition 2.3.1
S	sample; noise-free: p. 33, Definition 3.1.1; noisy: p. 36, Definition 3.2.1
$\mathrm{supp}(\hat{f})$	Fourier support of f; p. 21, Definition 2.2.1
$\mathcal{T}, \mathcal{T}_\epsilon$	set of Fourier coefficients used for LMN-style learning; p. 48
T_P	noise operator corresponding to noise distribution P;

	p. 42, Definition 3.4.1
$\mathrm{thr}_D(f)$	Fourier threshold of f w.r.t. D, i.e., minimal absolute value of first-level Fourier coefficients; p. 92, Equation (5.2)
U_n	uniform distribution on $\{0,1\}^n$; p. 18
$\mathrm{Var}[X]$	variance of random variable X
$V(n,d)$	volume of Hamming ball of radius d in $\{0,1\}^n$; p. 17, Equation (2.1)
x	attribute vector $x \in \{0,1\}^n$, distributed according to D
x_I	expanded variable with value $\bigoplus_{i\in I} x_i$; p. 22, Definition 2.2.2
$x\vert_I$	restriction of x to I; p. 17
x^k	attribute vector of k^{th} example
y	mostly: classification $y = f(x)$
y^k	classification of k^{th} example
$\mathbb{Z}$	set of integers
α_I	value related to $\vert E_I\vert$: $\vert E_I\vert \approx \alpha_I m^2$ for noise-free data; p. 61, Equation (4.3)
α_I^{ab}	$\Pr[x_I = a \wedge y = b]$ in a noise-free example; p. 60, Equation (4.2)
β_I	value related to $\vert E_I\vert$: $\vert E_I\vert \approx \alpha_I m^2$ for noisy data; p. 73, Equation (4.12)
β_I^{ab}	$\Pr[x_I = a \wedge y = b]$ in a noisy example; p. 73, Equation (4.11)
γ_a	lower bound for $\vert 1-2p_i\vert$; p. 39, Definition 3.2.6
γ_b	lower bound for $\vert 1-2\eta\vert$
γ_c	lower bound for $\vert d_i\vert$ and $\vert 1-d_i\vert$; p. 89
δ	confidence parameter, p. 36, Definition 3.2.2
$\Delta_P(f,g)$	noisy distance between concepts f and g; p. 44, Equation (3.9)
$\Delta_P^{\epsilon}(\mathcal{C})$	minimum noisy distance between ϵ-far concepts in $\mathcal{C}$; p. 44, Equation (3.10)
$\Gamma_P(\mathcal{C})$	noise stability of class $\mathcal{C}$ under noise distribution P; p. 46, Definition 3.4.6
$\Gamma_P(f)$	noise stability of function f under noise distribution P; p. 46, Definition 3.4.6
ϵ	mostly: accuracy parameter; p. 36, Definition 3.2.2
ζ	classification noise bit $\zeta \in \Omega$; p. 36, Definition 3.2.1
ζ^k	classification noise bit of k^{th} example
η	classification noise rate; p. 36
θ	lower bound for $\vert 1-2d_i\vert$; p. 94, Lemma 5.6.3
λ	degree of success/failure for the greedy algorithm; p. 58, Definition 4.1.4
λ_I	p. 40, (3.6); also: eigenvalue of T_P; p. 42, Lemma 3.4.2
ξ	attribute noise vector $\xi \in \{0,1\}^n$, distributed according to P;

	p. 36, Definition 3.2.1
ξ^k	attribute noise vector of k^{th} example
σ_i	standard deviation of $x_i \in \{0,1\}$ w.r.t. attribute distribution D; p. 19
τ	parameter for the level of lowness/Fourier-accessibility; p. 26, Definition 2.4.1 and p. 31, Definition 2.4.9
χ_I	$\{-1,+1\}$-parity of bits indexed by I; p. 19, Equation (2.5)
Ω	Boolean range, either $\Omega = \{0,1\}$ or $\Omega = \{-1,+1\}$; p. 15
$[n]$	$[n] = \{1,\ldots,n\}$
$\oplus$	exclusive or (= sum modulo 2); p. 16
$\triangle$	symmetric difference of sets; p. 16
$\sim$	distributed according to
$\langle\cdot\,,\cdot\rangle_D$	inner product on $\mathbb{R}^{\{0,1\}^n}$ induced by distribution D on $\{0,1\}^n$; p.18, Equation (2.2)
$\langle f \mid f \in \mathcal{C}\rangle$	real linear span of functions $f \in \mathcal{C} \subseteq \mathbb{R}^{\{0,1\}^n}$
$\lVert\cdot\rVert_D$	norm induced by $\langle\cdot\,,\cdot\rangle_D$; p. 18, Equation (2.3)
$\lVert\cdot\rVert_p$	p-norm ($p \geq 1$) induced by distribution D on $\{0,1\}^n$; p. 20, Equation (2.9)

Bibliography

[AB96] Tatsuya Akutsu and Feng Bao. Approximating Minimum Keys and Optimal Substructure Screens. In Jin-yi Cai and C. K. Wong, editors, *Computing and Combinatorics, Second Annual International Conference, COCOON '96, Hong Kong, June 17-19, 1996, Proceedings*, volume 1090 of *Lecture Notes in Comput. Sci.*, pages 290–299. Springer, 1996.

[ABSS97] Sanjeev Arora, Lászlő Babai, Jacques Stern, and Z Sweedyk. The Hardness of Approximate Optima in Lattices, Codes, and Systems of Linear Equations. *J. Comput. System Sci.*, 54(2):317–331, 1997.

[AD94] Hussein Almuallim and Thomas G. Dietterich. Learning Boolean Concepts in the Presence of Many Irrelevant Features. *Artificial Intelligence*, 69(1-2):279–305, September 1994.

[AIS93] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami. Mining Association Rules between Sets of Items in Large Databases. In Peter Buneman and Sushil Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., May 26-28, 1993*, pages 207–216. ACM Press, 1993.

[AL88] Dana Angluin and Philip D. Laird. Learning From Noisy Examples. *Machine Learning*, 2(4):343–370, April 1988.

[AMK00] Tatsuya Akutsu, Satoru Miyano, and Satoru 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, October 2000.

[AMK03] Tatsuya Akutsu, Satoru Miyano, and Satoru Kuhara. A Simple Greedy Algorithm for Finding Functional Relations: Efficient Implementation and Average Case Analysis. *Theoret. Comput. Sci.*, 292(2):481–495, January 2003.

[Ang87] Dana Angluin. Queries and Concept Learning. *Machine Learning*, 2(4):319–342, 1987.

[AR03] Jan Arpe and Rüdiger Reischuk. Robust Inference of Relevant Attributes. In Ricard Gavaldà, Klaus P. Jantke, and Eiji Takimoto, editors, *Algorithmic Learning Theory, 14th International Conference, ALT 2003, Sapporo, Japan, October 2003, Proceedings*, volume 2842 of *Lecture Notes in Artificial Intelligence*, pages 99–113. Springer, 2003.

[AR06] Jan Arpe and Rüdiger Reischuk. Learning Juntas in the Presence of Noise. In Jin-Yi Cai, S. Barry Cooper, and Angsheng Li, editors, *Theory and Applications of Models of Computation, Third Annual Conference, TAMC 2006, Beijing, China, May 2006, Proceedings*, volume 3959 of *Lecture Notes in Comput. Sci.*, pages 387–398, 2006. Invited to appear in special issue of TAMC 2006 in *Theoret. Comput. Sci., Series A*.

[Art91] Michael Artin. *Algebra.* Prentice Hall, 1991.

[AS92] Noga Alon and Joel Spencer. *The Probabilistic Method.* Wiley-Intersci. Ser. Discrete Math. Optim. John Wiley and Sons, 1992.

[Bah61] Raghu Raj Bahadur. A Representation of the Joint Distribution of Responses to n Dichotomous Items. In Herbert Solomon, editor, *Studies in Item Analysis and Prediction*, pages 158–168. Stanford University Press, Stanford, California, 1961.

[BCJ93] Avrim Blum, Prasad Chalasani, and Jeffrey C. Jackson. On Learning Embedded Symmetric Concepts. In Leslie G. Valiant and Manfred K. Warmuth, editors, *Proceedings of the Sixth Annual ACM Conference on Computational Learning Theory (COLT 1993), July 26-28, 1993, Santa Cruz, CA, USA*, pages 337–346. ACM, 1993.

[Bec75] William Beckner. Inequalities in Fourier Analysis. *Ann. of Math. (2)*, 102(1):159–182, July 1975.

[BEHW87] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Occam's Razor. *Inform. Process. Lett.*, 24(6):377–380, April 1987.

[BEHW89] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis Dimension. *J. ACM*, 36(4):929–965, October 1989.

[Ber98] Anna Bernasconi. *Mathematical Techniques for the Analysis of Boolean Functions.* PhD thesis, Università degli Studi di Pisa, Dipartimento di Ricerca in Informatica, March 1998.

[BFJ$^+$94] Avrim Blum, Merrick Furst, Jeffrey C. Jackson, Michael Kearns, Yishay Mansour, and Steven Rudich. Weakly Learning DNF and Characterizing Statistical Query Learning Using Fourier Analysis. In *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing, 23-25 May 1994, Montréal, Québec, Canada*, pages 253–262, 1994.

[BHI$^+$03] Endre Boros, Takashi Horiyama, Toshihide Ibaraki, Kazuhisa Makino, and Mutsunori Yagiura. Finding Essential Attributes from Binary Data. *Ann. Math. Artif. Intell.*, 39(3):223–257, November 2003.

[BJT03] Nader H. Bshouty, Jeffrey C. Jackson, and Christino Tamon. Uniform-distribution attribute noise learnability. *Information and Computation*, 187(2):277–290, April 2003.

[BKS99] Itai Benjamini, Gil Kalai, and Oded Schramm. Noise Sensitivity of Boolean Functions and Applications to Percolation. *Inst. Hautes Études Sci. Publ. Math.*, 90:5–43, 1999.

[BKW03] Avrim Blum, Adam Kalai, and Hal Wasserman. Noise-Tolerant Learning, the Parity Problem, and the Statistical Query Model. *J. ACM*, 50(4):506–519, September 2003.

[BL97] Avrim Blum and Pat Langley. Selection of Relevant Features and Examples in Machine Learning. *Artificial Intelligence*, 97(1-2):245–271, December 1997.

[Blu03] Avrim Blum. Learning a Function of r Relevant Variables. In Bernhard Schölkopf and Manfred K. Warmuth, editors, *Computational Learning Theory and Kernel Machines, 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop,* *COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003, Proceedings*, volume 2777 of *Lecture Notes in Computer Science*, pages 731–733. Springer, 2003.

[BMvT78] Elwyn R. Berlekamp, Robert J. McEliece, and Henk C. A. van Tilborg. On the inherent intractability of certain coding problems. *IEEE Trans. Inform. Theory*, IT-24(3):384–386, 1978.

[Bon70] Aline Bonami. Étude des coefficients de Fourier des fonctions de $l^p(g)$. *Ann. Inst. Fourier*, 20(2):335–402, 1970.

[BT96] Nader H. Bshouty and Christino Tamon. On the Fourier Spectrum of Monotone Functions. *J. ACM*, 43(4):747–770, July 1996.

[Chv79] Vašek Chvátal. A Greedy Heuristic for the Set Covering Problem. *Math. Oper. Res.*, 4(3):233–235, 1979.

[DF99] Rodney G. Downey and Michael R. Fellows. *Parameterized Complexity.* Monographs in Computer Science. Springer, New York, 1999.

[DFVW99] Rodney G. Downey, Michael R. Fellows, Alexander Vardy, and Geoff P. Whittle. The Parameterized Complexity of Some Fundamental Problems in Coding Theory. *SIAM Journal on Computing*, 29(2):545–570, October–December 1999.

[DG95] Scott E. Decatur and 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.*, pages 353–360. ACM Press, 1995.

[Edm71] Jack Edmonds. Matroids and the Greedy Algorithm. *Math. Program.*, 1(1):127–136, December 1971.

[FA05] Daiji Fukagawa and Tatsuya Akutsu. Performance Analysis of a Greedy Algorithm for Inferring Boolean Functions. *Inform. Process. Lett.*, 93(1):7–12, January 2005.

[Fei98] Uriel Feige. A Threshold of $\ln n$ for Approximating Set Cover. *J. ACM*, 45(4):634–652, July 1998.

[Fel05] Vitaly Feldman. On Attribute Efficient and Non-adaptive Learning of Parities and DNF Expressions. In Peter Auer and Ron Meir, editors, *Learning Theory, 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, June 27-30, 2005, Proceedings*, volume 3559 of *Lecture Notes in Artificial Intelligence*, pages 576–590. Springer, 2005.

[FGKP06] Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. New Results for Learning Noisy Parities and Halfspaces. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), October 22-24, Berkeley, CA*, 2006. To appear.

[FJS91] Merrick L. Furst, Jeffrey C. Jackson, and Sean W. Smith. Improved Learning of AC^0 Functions. In Leslie G. Valiant and Manfred K. Warmuth, editors, *Proceedings of the Fourth Annual Workshop on Computational Learning Theory (COLT 1991), Santa Cruz, California, USA*, pages 317–325. Morgan Kaufmann, 1991.

[GJ79] Michael R. Garey and David S. Johnson. *Computers and Intractability—A Guide to the Theory of NP-Completeness.* Freeman, 1979.

[GS95] Sally A. Goldman and Robert H. Sloan. Can PAC Learning Algorithms Tolerate Random Attribute Noise? *Algorithmica*, 14(1):70–84, July 1995.

[Hau88] David Haussler. Quantifying Inductive Bias: AI Learning Algorithms and Valiant's Learning Framework. *Artif. Intell.*, 36(2):177–221, September 1988.

[Hoc82] Dorit S. Hochbaum. Approximation Algorithms for the Set Covering and Vertex Cover Problems. *SIAM Journal on Computing*, 11(3):555–556, August 1982.

[Hoe63] Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

[HSW92] David Helmbold, Robert Sloan, and Manfred K. Warmuth. Learning integer lattices. *SIAM Journal on Computing*, 21(2):240–266, April 1992.

[JKP94] George H. John, Ron Kohavi, and Karl Pfleger. Irrelevant Features and the Subset Selection Problem. In William W. Cohen and Haym Hirsh, editors, *Machine Learning, Proceedings of the Eleventh International Conference, Rutgers University, New Brunswick, NJ, USA, July 10-13, 1994*, pages 121–129. Morgan Kaufmann, 1994.

[Joh74] David S. Johnson. Approximation Algorithms for Combinatorial Problems. *J. Comput. System Sci.*, 9(3):256–278, December 1974.

[Kat04] Yitzhak Katznelson. *An Introduction to Harmonic Analysis*. Cambridge University Press, Cambridge, England, third edition, 2004.

[Kea98] Michael Kearns. Efficient Noise-Tolerant Learning from Statistical Queries. *J. ACM*, 45(6):983–1006, November 1998.

[KKL88] Jeff Kahn, Gil Kalai, and Nathan Linial. The Influence of Variables on Boolean Functions (extended abstract). In *29th Annual Symposium on Foundations of Computer Science (FOCS '88)*, pages 68–80, White Plains, New York, October 1988. IEEE Computer Society.

[KL93] Michael Kearns and Ming Li. Learning in the Presence of Malicious Errors. *SIAM Journal on Computing*, 22(4):807–837, August 1993.

[KL06] Johannes Köbler and Wolfgang Lindner. Learning Boolean Functions under the uniform distribution via the Fourier Transform. *Bulletin of the EATCS*, 89:48–78, June 2006.

[KM93] Eyal Kushilevitz and Yishay Mansour. Learning Decision Trees Using the Fourier Spectrum. *SIAM Journal on Computing*, 22(6):1331–1348, December 1993.

[KMM05] Mihail N. Kolountzakis, Evangelios Markakis, and 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`.

[KS06] Adam R. Klivans and Rocco A. Servedio. Toward Attribute Efficient Learning of Decision Lists and Parities. *J. Mach. Learn. Res.*, 7:587–602, April 2006.

[KT05] Jon Kleinberg and Éva Tardos. *Algorithm Design*. Addison Wesley, 2005.

[Lan93] Serge Lang. *Algebra*. Addison-Wesley, 3rd edition, 1993.

[Lec71] Robert J. Lechner. Harmonic Analysis of Switching Functions. In Amar Mukhopadhyay, editor, *Recent Developments in Switching Theory*, pages 121–228. Academic Press, 1971.

[Lev86] Leonid A. Levin. Average Case Complete Problems. *SIAM Journal on Computing*, 15(1):285–286, February 1986.

[Lit87] Nick Littlestone. Learning Quickly When Irrelevant Attributes Abound: A New Linear-threshold Algorithm. *Machine Learning*, 2(4):285–318, 1987.

[Lit89] Nick Littlestone. From on-line to batch learning. In *Proceedings of the Second Annual Workshop on Computational Learning Theory, COLT 1989, July 31 - August 2, 1989, Santa Cruz, CA, USA*, pages 269–284. Morgan Kaufmann, 1989.

[LMMV05] Richard J. Lipton, Evangelos Markakis, Aranyak Mehta, and Nisheeth K. Vishnoi. On the Fourier Spectrum of Symmetric Boolean Functions with Applications to Learning Symmetric Juntas. In *20th Annual IEEE Conference on Computational Complexity (CCC '05)*, pages 112–119, 2005.

[LMN93] Nathan Linial, Yishay Mansour, and Noam Nisan. Constant Depth Circuits, Fourier Transform, and Learnability. *J. ACM*, 40(3):607–620, July 1993.

[Loo53] Lynn H. Loomis. *An Introduction to Abstract Harmonic Analysis.* The University Series in Higher Mathematics. D. van Nostrand Company, Inc., Princeton, New Jersey, 1953.

[Man94] Yishay Mansour. Learning Boolean Functions via the Fourier Transform. In V.P. Roychodhury, K.-Y. Siu, and A. Orlitsky, editors, *Theoretical Advances in Neural Computation and Learning*, pages 391–424. Kluwer Academics, 1994.

[MO03] Elchanan Mossel and Ryan W. O'Donnell. On the Noise Sensitivity of Monotone Functions. *Random Structures Algorithms*, 23(3):333–350, October 2003.

[Mon81] Gaspard Monge. Déblai et Remblai. *Mémoires de l'Académie des Sciences*, pages 666–704, 1781.

[MOS04] Elchanan Mossel, Ryan W. O'Donnell, and Rocco A. Servedio. Learning functions of k relevant variables. *J. Comput. System Sci.*, 69(3):421–434, November 2004.

[MR92] Heikki Mannila and Kari-Jouko Räihä. On the Complexity of Inferring Functional Dependencies. *Discrete Appl. Math.*, 40(2):237–243, 1992.

[MTT04] Akinobu Miyata, Jun Tarui, and Etsuji Tomita. Learning Boolean Functions in AC^0 on Attribute and Classification Noise. In Shai Ben-David, John Case, and Akira Maruoka, editors, *Algorithmic Learning Theory, 15th International Conference, ALT 2004, Padova, Italy, October 2-5, 2004, Proceedings*, volume 3244 of *Lecture Notes in Artificial Intelligence*, pages 142–155. Springer, 2004.

[O'D03] Ryan W. O'Donnell. *Computational Applications Of Noise Sensitivity.* PhD thesis, Department of Mathematics, Massachusetts Institute of Technology, June 2003.

[PV88] Leonard Pitt and Leslie G. Valiant. Computational Limitations on Learning from Examples. *J. ACM*, 35(4):965–984, October 1988.

[Rad42] Richard Rado. A Theorem on Independence Relations. *Q. J. Math.*, 13:83–89, 1942.

[RHRP05] Bernard Rosell, Lisa Hellerstein, Soumya Ray, and David Page. Why Skewing Works: Learning Difficult Boolean Functions with Greedy Tree Learners. In Luc De Raedt and Stefan Wrobel, editors, *Proceedings of the 22nd International Conference on Machine Learning. Bonn, Germany.*, volume 119 of *ACM International Conference Proceeding Series*, pages 728–735, 2005.

[SD90] Ron Shamir and Brenda Dietrich. Characterization and Algorithms for Greedily Solvable Transportation Problems. In *Proceedings of the First Annual ACM-SIAM Symposium on Discrete Algorithms, 22-24 January 1990, San Francisco, California*, pages 358–366. ACM/SIAM, 1990.

[Ser04] Rocco A. Servedio. On learning monotone DNF under product distributions. *Inform. and Comput.*, 193(1):57–74, August 2004.

[Ser05] Rocco A. Servedio. On learning embedded midbit functions. *Theoretical Computer Science*, 350(1):13–23, January 2005.

[Sla96] Petr Slavík. A Tight Analysis of the Greedy Algorithm for Set Cover. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on the Theory of Computing, Philadelphia, Pennsylvania, USA, May 22-24, 1996*, pages 435–441. ACM, 1996.

[Šte00] Daniel Štefankovič. Fourier Transforms in Computer Science. Master's thesis, Department of Computer Science, University of Chicago, October 2000.

[SV88] George Shackelford and 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*, pages 97–103. Morgan Kaufmann, 1988.

[Ter99] Audrey Terras. *Fourier Analysis on Finite Groups and Applications.* Cambridge Univ. Press, Cambridge, U.K., 1999.

[Tur93] György Turán. Lower Bounds for PAC Learning with Queries. In Leslie G. Valiant and Manfred K. Warmuth, editors, *Proceedings of the Sixth Annual ACM Conference on Computational Learning Theory (COLT 1993), July 26-28, 1993, Santa Cruz, CA, USA*, pages 384–391. ACM, 1993.

[UTW97] Ryuhei Uehara, Kensei Tsuchida, and Ingo Wegener. Optimal Attribute-Efficient Learning of Disjunction, Parity and Threshold Functions. In Shai Ben-David, editor, *Computational Learning Theory, Third European Conference, EuroCOLT '97, Jerusalem, Israel, March 17-19, 1997, Proceedings*, volume 1208 of *Lecture Notes in Comput. Sci.*, pages 171–184. Springer, 1997.

[Val84] Leslie G. Valiant. A Theory of the Learnable. *Commun. ACM*, 27(11):1134–1142, November 1984.

[Val99] Leslie G. Valiant. Projection Learning. *Machine Learning*, 37(2):115–130, November 1999.

[Vin02] A. Vince. A Framework for the Greedy Algorithm. *Discrete Appl. Math.*, 121(1-3):247–260, September 2002.

Curriculum Vitae

Personal Data

Name	Jan Arpe
Date of birth	June 16, 1977
Place of birth	Kiel, Germany
Family status	married, two daughters (3 years and 1 year old)

Education and Work Experience

1983 – 1984	Grundschule Himmelpforten (primary school)
1984 – 1987	Nikolaus-Groß-Grundschule Lünen (primary school)
1987 – 1996	Gymnasium Nieder-Olm (secondary school)
1996 – 1999	Study of Mathematics (Mathematik) with minor subject Computer Science (Informatik) at the Johannes-Gutenberg-Universität Mainz 10/1998 Vordiplom (pre-diploma)
1999 – 2002	Study of Mathematics (Mathematik) with minor subject Computer Science (Informatik) at the Ludwig-Maximilians-Universität München 02/2002 Diplom (diploma), Diplomarbeit (diploma thesis): "Berechnung sekundärer Koeffizientengruppen des $SO(3) \times S^1$-äquivarianten Abbildungsgrades" ("Computation of secondary coefficient groups of the $SO(3) \times S^1$-equivariant mapping degree")
since 03/2002	Research associate at the Institut für Theoretische Informatik of the Universität zu Lübeck