Electronic Colloquium on Computational Complexity, Report No. 65 (2006)

When Does Greedy Learning of Relevant Features Succeed? — A Fourier-based Characterization —

Jan Arpe* Rüdiger Reischuk

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

May 12, 2006

Abstract

Detecting the relevant attributes of an unknown target concept is an important and well studied problem in algorithmic learning. Simple greedy strategies have been proposed that seem to perform reasonably well in practice if a sufficiently large random subset of examples of the target concept is provided. Introducing a new notion called *Fourier-accessibility* allows us to characterize the class of Boolean functions precisely for which a standard greedy learning algorithm successfully learns all relevant attributes. Technically, this is achieved by deriving new relations between the learnability of a function and its Fourier spectrum. We prove that if the target concept is Fourier-accessible, then the success probability of the greedy algorithm can be made arbitrarily close to one. On the other hand, if the target concept is *not* Fourier-accessible, then the error probability tends to one.

Keywords: greedy algorithms, learning relevant features, Fourier analysis of Boolean functions

1 Introduction

For many application areas, greedy strategies are natural, important, 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 [19, Chapter 4]).

For the vast majority of optimization problems, however, greedy heuristics do not always achieve optimal solutions. In such cases, the behaviour of greedy algorithms is hardly understood. In particular, 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 [23].

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 [18, 14, 24]), which has been proven to be best possible by Feige [15].

Confronted with an unknown *target concept* $f : A^n \to B$ on n variables, the problem of detecting which variables x_i (also referred to as *attributes* or *features*) are relevant to f is known as *relevant feature selection*. This problem lies at the heart of many data mining applications; this is

*Supported by DFG research grant Re 672/4.

ISSN 1433-8092

particularly the case if f depends only on a small number d of all n attributes—such concepts are called *d-juntas*. A survey of this topic has been provided by Blum and Langley [11].

To infer relevant attributes from a randomly drawn sample $S = (x^k, y^k)_{k=1,\ldots,m}$ with $x^k = (x_1^k, \ldots, x_n^k) \in A^n$ and $y^k = f(x^k) \in B$, 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 (i.e., $h(x^k) = y^k$ for all k) that depends only on the variables in R. By standard arguments [12], once the sample size m exceeds $\text{poly}(2^d, \log n)$, with high probability there remains only one such hypothesis—the target concept itself. Finding such a set R is equivalent to solving the following SET COVER instance. The ground set is the set of all pairs $\{k, \ell\}$ such that $y^k \neq y^\ell$. 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 [18] or Chvatal [14].

For relevant feature selection, this approach has been proposed by Almuallin and Dietterich [4] and Akutsu and Bao [1]. Experimental results have been obtained for artificially generated instances as well as for real-world data from various areas, see Almuallim and Dietterich [4], Akutsu, Miyano, and Kuhara [2, 3], and Boros et al. [13]. Akutsu et al. [3] have shown how to implement GREEDY in such a way that its running time is only $O(mn)$.

In this paper, we are mainly concerned with Boolean concepts f, i.e., $A = B = \{0, 1\}$, and uniformly distributed attributes. For this case, Akutsu et al. [3] 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., monomials) and that a small sample of size $\text{poly}(2^d, \log n)$already suffices. Fukagawa and Akutsu [16] 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).

The main result of this paper is a concise characterization of the class of target concepts for which GREEDY is able to infer the relevant variables. This class properly contains the concept classes mentioned above. The new characterization is based on a property of the Fourier spectrum of the target concept, which we call *Fourier-accessibility*. A function $f : \{0, 1\}^n \to \{0, 1\}$ is *Fourier-accessible* if for each relevant variable, one can find a sequence $\emptyset \subsetneq I_1 \subsetneq \ldots \subsetneq I_s \subseteq [n]$ such that $i \in I_s$ and for all $j \in \{1, \ldots, s\}$, $|I_j \setminus I_{j-1}| = 1$ and $\hat{f}(I_j) \neq 0$. Equivalently, $f \not\equiv 0$ is Fourier-accessible if and only if for every x_i that is relevant to f, there exists $I \subseteq [n]$ with $i \in I$ and a path from $\emptyset$ to I in the *Fourier support graph*. This graph is the subgraph of the n-dimensional Hamming cube induced by the *Fourier support* of f, i.e., the set of subsets I of $[n]$ such that $\hat{f}(I) \neq 0$.

We prove that GREEDY correctly infers all relevant variables of Fourier-accessible d-juntas $f : \{0, 1\}^n \to \{0, 1\}$ under the uniform distribution from $m = \text{poly}(2^d, \log n, \log(1/\delta))$ examples with probability at least $1 - \delta$. On the other hand, we show that if a function f is *not* Fourier-accessible, then the error probability of GREEDY is at least $1 - d^2/(n - d)$. In particular, this probability tends to 1 as d is fixed and $n \to \infty$ or as $d \to \infty$ and $n \in \omega(d^2)$.

In a previous paper [6], we have considered an alternative greedy strategy, called GREEDY RANKING. It ranks the sets of the set cover instance by their size and then successively picks the largest ones (without deleting anything) until all elements are covered. We have shown that GREEDY RANKING successfully infers all relevant variables of a function f if a certain *gap condition* $\Delta_i(f) > 0$ is satisfied for each relevant variable x_i. The *gap* $\Delta_i(f)$ measures how much the expected number of pairs covered by x_i differs from the expected number of pairs covered by some irrelevant variable. In this paper, we show that the gap condition is equivalent to the statement that for all relevant variables x_i, $\hat{f}(i) \neq 0$. This, in turn, is equivalent to Fukagawa and Akutsu's condition of f being unbalanced with respect to all of its relevant attributes. We call functions that satisfy

this property 1-*low*. Prominent examples of 1-low functions are monotone functions, monomials and clauses of arbitrary literals, and random functions (see Blum and Langley [11] and Mossel et al. [22] for the latter item).

In general, for $t \in \{1, \ldots, n\}$, a function $f : \{0,1\}^n \to \{0,1\}$ is said to be *t-low* (see also [7]) if for each relevant variable x_i of f, there exists $I \subseteq [n]$ with $i \in I$ such that $|I| \leq t$ and $\hat{f}(I) \neq 0$.

Similarly to the analysis of GREEDY, we obtain that GREEDY RANKING learns *exactly* the class of 1-low functions. Since all 1-low functions are Fourier-accessible (but not vice versa), these results also provide an analytic argument that the dynamic variant GREEDY (that adjusts the covering sets after each round) is in general preferable to the static version GREEDY RANKING.

There is a long tradition of relating algorithmic learning problems to spectral properties of Boolean functions, see, e.g., Linial, Mansour, and Nisan [20], Mansour [21], Blum et al. [10]. Specifically, Mossel, O'Donnell, and Servedio [22] have combined spectral and algebraic methods to reduce the worst-case running time for learning the class of all n-ary d-juntas to roughly $n^{0.7 \cdot d}$ (a trivial approach is to test all $\Theta(n^d)$ sets of potentially relevant variables).

Here we extend this kind of approach to derive our new results. The GREEDY algorithm investigated in this paper does *not* exploit any properties of the Fourier spectrum explicitly. However, 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.

This paper is organized as follows. The terminology and the learning model are introduced in Section 2. The reduction to SET COVER and the GREEDY algorithm are presented in Section 3. Section 4 provides two major lemmas used in the proof of our main results for GREEDY, which are presented in Section 5. Some variations of GREEDY are discussed in Section 6.1. Further results for generalizations to functions $f : \mathbb{Z}_r^n \to \{0,1\}$ and non-uniform distributions are given in Section 6.2. In Section 6.3, a few more open problems are posed.

2 Preliminaries

For $n \in \mathbb{N}$, let $[n] = \{1, \ldots, n\}$. We consider the problem of inferring the relevant variables of an unknown function $f : \{0,1\}^n \to \{0,1\}$ from randomly drawn examples. Such a function f will also be called a *concept*. Variable x_i is *relevant* to f if $f_{x_i=0} \neq f_{x_i=1}$, where $f_{x_i=a}$ denotes the restriction of f with variable x_i set to a. The set of variables that are relevant to f is denoted by $\mathrm{rel}(f)$, whereas $\mathrm{irrel}(f)$ denotes the set of variables that are *irrelevant* (i.e., not relevant) to f. A function with $|\mathrm{rel}(f)| \leq d$ is called a *d-junta*. Let $D : \{0,1\}^n \to [0,1]$ be a probability distribution. We assume that an algorithm for inferring the relevant variables receives a sequence S of randomly generated *examples* (x^k, y^k), $k \in [m]$, where $x^k \in \{0,1\}^n$ is drawn according to D and $y^k = f(x^k) \in \{0,1\}$. Such a sequence is called a *sample* for f of size m. If for another concept h, $y^k = h(x^k)$ for all $k \in [m]$, h is said to be *consistent* with S.

A concept $f : \{0,1\}^n \to \{0,1\}$ may also be considered as a Bernoulli random variable, thus we will use the notation $\Pr[f = b] = \Pr_{x \sim D}[f(x) = b]$ for $b \in \{0,1\}$ and $\mathrm{Var}(f) = \Pr[f = 0] \Pr[f = 1]$.

We will mostly be concerned with the case that D is the uniform distribution on $\{0,1\}^n$. Extensions to $f : \mathbb{Z}_r^n \to \{0,1\}$ and to non-uniform attribute distributions D are discussed in Section 6.2. The following well-known Chernoff bound can, e.g., be found in Alon and Spencer [5].

Fact 2.1 (Chernoff Bound). Let X be a random variable that is binomially distributed with parameters n and p, and let $\mu = pn$ be the expectation of X. For all ϵ with $0 \leq \epsilon \leq 1$, $\Pr[|X - \mu| > \epsilon n] < 2e^{-2\epsilon^2 n}$.

Consider the space $\mathbb{R}^{\{0,1\}^n}$ of real-valued functions on the hypercube. The inner product $\langle f, g\rangle = 2^{-n}\sum_{x\in\{0,1\}^n} f(x)g(x)$ turns $\mathbb{R}^{\{0,1\}^n}$ into a Hilbert space of dimension 2^n with orthonormal basis $(\chi_I \mid I \subseteq [n])$, where $\chi_I(x) = (-1)^{\sum_{i\in I} x_i}$ for $x \in \{0,1\}^n$, see for example Bernasconi [9].

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

$$\hat{f}(I) = 2^{-n} \sum_{x\in\{0,1\}^n} f(x) \cdot \chi_I(x) \,.$$

If $I = \{i\}$, we write $\hat{f}(i)$ instead of $\hat{f}(\{i\})$. We have the *Fourier expansion* $f(x) = \sum_{I\subseteq[n]} \hat{f}(I)\cdot\chi_I(x)$ for all $x \in \{0,1\}^n$.

Definition 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 $\mathrm{supp}(\hat{f})$.

Fourier coefficients are connected to relevant variables as follows (cf. [22, 7]):

Lemma 2.1. *Let $f : \{0,1\}^n \to \{0,1\}$. Then for all $i \in [n]$, x_i is relevant to f if and only if there exists $I \subseteq [n]$ such that $i \in I$ and $\hat{f}(I) \neq 0$.*

Hence whenever we find a nonzero Fourier coefficient $\hat{f}(I)$, we know that all variables x_i, $i \in I$, are relevant to f. Moreover, all relevant variables can be detected in this way, and we only have to check out subsets of size at most $d = |\mathrm{rel}(f)|$. However, there are $\Theta(n^d)$ such subsets, an amount that one would generally like to reduce. For the class of all n-ary d-juntas, the best known learning algorithm to date that for *all* concepts is guaranteed to find the relevant attributes is due to Mossel et al. [22] and runs in time roughly $n^{0.7\cdot d}$. As discussed above, greedy heuristics require only time polynomial in n with an exponent independent of d.

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

Definition 2.2 (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 $\emptyset \subsetneq I_1 \subsetneq \ldots \subsetneq I_s \subseteq [n]$ 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 called *Fourier-accessible* if and only if every variable that is relevant to f is also accessible, i.e., $\mathrm{acc}(f) = \mathrm{rel}(f)$.

Equivalently, x_i is accessible if and only if there exists $I \in \mathrm{supp}(\hat{f})$ with $i \in I$ such that there is a path in $\mathrm{FSG}(f)$ from $\emptyset$ to I. Since $\hat{f}(\emptyset) = \Pr[f(x) = 1]$, $\emptyset \in \mathrm{supp}(\hat{f})$ whenever $f \not\equiv 0$. Hence f is Fourier-accessible if and only if the union of all subsets $I \in \mathrm{supp}(\hat{f})$ that belong to the connected component of $\emptyset$ in $\mathrm{FSG}(f)$ equals $\mathrm{rel}(f)$.

Throughout the paper, if f is clear from the context, we call a variable that is relevant to f simply *relevant*. Similarly, a variable that is accessible with respect to f is simply called *accessible*.

Simple examples of a Fourier-accessible function f_1 and a non-Fourier-accessible function f_2 are given in Table 1. The corresponding Fourier support graphs are presented in Figure 1.

For $R \subseteq [n]$ and $v \in \{0,1\}^R$, denote by $f_{R\leftarrow v}$ the restriction of f, in which the variables x_i, $i \in R$, are set to the values provided by v. The following lemma reveals a connection between vanishing Fourier coefficients of functions and their restrictions.

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

Table 1: Examples of Boolean functions and their Fourier spectra.

(a) FSG(f_1). (b) FSG(f_2).

Figure 1: Fourier support graphs of functions f_1 and f_2 presented in Table 1.

Lemma 2.2. *Let $f : \{0,1\}^n \to \{0,1\}$, $R \subsetneq [n]$, and $i \in [n] \setminus R$. If for all $I \subseteq R$, $\hat{f}(I \cup \{i\}) = 0$, then for all $v \in \{0,1\}^R$, $\widehat{f_{R \leftarrow v}}(i) = 0$.*

Proof. Consider the real multi-linear polynomial p with variables $z_i = (-1)^{x_i}$ that represents f, i.e., $p(z_1, \ldots, z_n) = f(x_1, \ldots, x_n)$ for all $x \in \{0,1\}^n$. The coefficient of $\prod_{i \in I} z_i$ in p is $\hat{f}(I)$. Since all coefficients $\hat{f}(I \cup \{i\})$, $I \subseteq R$, vanish in p, the coefficient $\hat{f}(i)$ must also vanish in the polynomial $\tilde{p}$ that represents $f_{R \leftarrow v}$. Hence $\widehat{f_{R \leftarrow v}}(i) = 0$. □

Using a standard result due to Blumer et al. [12], Almuallim and Dietterich [4] proved that if the sample size is sufficiently large, then the target concept is the only concept that depends on at most d variables and that is consistent with the sample:

Lemma 2.3 ([4]). *Let $f : \{0,1\}^n \to \{0,1\}$, $d = |\operatorname{rel}(f)|$, and $\delta > 0$. Let S be a uniformly distributed sample of size*

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

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

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

3 The Reduction to Set Cover and the Greedy Algorithm

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 (see also [3, 6]). 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 $[m] = A^{(0)} \cup A^{(1)}$. Given S, our primary goal is to determine a set of variables $R \subseteq \{x_1, \ldots, x_n\}$ 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 g may not be identical to the original concept f, nor may the set R contain all relevant variables of f.

GREEDY

```
1: input S = ((x_1^k, ..., x_n^k), y^k)_{k∈[m]}
2: E ← {{k, ℓ} | k, ℓ ∈ [m], y^k ≠ y^ℓ}
3: R ← ∅
4: while E ≠ ∅ do
5:   for i = 1 to n do
6:     E_i ← {{k, ℓ} ∈ E | x_i^k ≠ x_i^ℓ}
7:   select i ∈ [n] with maximum |E_i|
8:   E ← E \ E_i
9:   R ← R ∪ {x_i}
10: output GREEDY(S) = R
```

Figure 2: Algorithm GREEDY.

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 $\{k, \ell\} \in E$ may be *covered* by attribute x_i iff $x_i^k \neq x_i^\ell$. The set of edges that may be covered by x_i is denoted by E_i. The characteristic vector of an edge $e = \{k, \ell\} \in E$ is

$$c(e) = (c_1(e), \ldots, c_n(e)) = (x_1^k \oplus x_1^\ell, \ldots, x_n^k \oplus x_n^\ell) \ . \tag{2}$$

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 [4].

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

Lemma 3.1. *Let $S \in (\{0,1\}^n \times \{0,1\})^m$ be a sample and $R \subseteq \{x_1, \ldots, x_n\}$. Then R explains S if and only if $E = \cup_{x_i \in R} E_i$, where E is the edge set of the functional relations graph G_S.*

The lemma provides a reduction from the problem of inferring small sets 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 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 in Figure 2.

If there are several sets of maximum cardinality in step 7 of GREEDY, it picks one of them at random. The notion of success for GREEDY is captured as follows.

Definition 3.1 (λ-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 $|\text{GREEDY}(S)| \leq \lambda \cdot |\text{rel}(f)|$ and $\text{GREEDY}(S) \supseteq \text{rel}(f)$. GREEDY is *successful* (or *succeeds*) if and only if it is 1-*successful*, i.e., $\text{GREEDY}(S) = \text{rel}(f)$, otherwise we say that it fails. GREEDY *λ-fails* if and only if it is not λ-successful.

4 Two Crucial Lemmas

In this section, we provide two key lemmas that will be used in the proofs of our main results in Section 5. For technical reasons, it proves useful to consider the *expanded attribute space* of

attributes

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

and the corresponding edge sets

$$E_I = \{\{k, \ell\} \in E \mid x_I^k \neq x_I^\ell\} \ .$$

Lemma 4.1 shows how to express the cardinality of $E_i \setminus (E_{i_1} \cup \ldots \cup E_{i_s})$ in terms of the cardinalities of the sets E_I, $I \subseteq \{i_1, \ldots, i_s\}$. Lemma 4.2 then shows how to estimate the cardinalities of the latter sets.

Lemma 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 $i^* \in [n] \setminus R$ 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 (2) of the characteristic vector $c(e) \in \{0,1\}^n$ of an edge $e \in E$. 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 $c_{I \cup \{i^*\}}(e) = c_I(e)$,

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

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

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

Now for all $j \in R$ and all $e \in E_j$, we have $e \in E_I$ for exactly half of the sets $I \subseteq R$, so that $\sum_{I \subseteq R} (c_{I \cup \{i^*\}}(e) - c_I(e)) = 0$ in this case. Consequently,

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

On the other hand, if $e \in E_{i^*} \setminus E'$, then $e \notin E_I$ for all $I \subseteq R$. Hence, $\sum_{I \subseteq R} (c_{I \cup \{i^*\}}(e) - c_I(e)) = 2^{|R|}$ in this case and thus

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

proving the claim. □

Before stating the second lemma, let us briefly take a closer look at the cardinalities $|E_i|$ for irrelevant variables x_i. Given examples (x^k, y^k) and (x^ℓ, y^ℓ), $k \neq \ell$, the probability that $\{k, \ell\} \in E_i$ is $\Pr[x_i^k \neq x_i^\ell \wedge y^k \neq y^\ell]$. Since for irrelevant x_i, the value of x_i^k is independent of the classification y^k, the probability equals

$$\Pr[x_i^k \neq x_i^\ell] \cdot \Pr[y^k \neq y^\ell] = \tfrac{1}{2} \cdot 2 \Pr[f = 0] \Pr[f = 1] = \mathrm{Var}(f) \ .$$

Hence the expectation of $|E_i|$ is $\frac{1}{2} \mathrm{Var}(f) m(m-1) \approx \frac{1}{2} \mathrm{Var}(f) m^2$ since there are $\frac{1}{2} m(m-1)$ pairs $\{k, \ell\}$ with $k, \ell \in [m]$ and $k \neq \ell$. A moment's reflection shows that also for $I \subseteq [n]$ with $I \not\subseteq \mathrm{rel}(f)$, the expectation of $|E_I|$ equals $\mathrm{Var}(f) m(m-1)/2$.

The following lemma generalizes this result to *arbitrary* $I \subseteq [n]$, revealing an unexpected relationship between the cardinalities $|E_I|$ and the Fourier coefficients $\hat{f}(I)$. This connection is spelled out in combination with a Chernoff style mass concentration of the cardinalities $|E_I|$. Note that for $I \subseteq [n]$ with $I \not\subseteq \mathrm{rel}(f)$, $\hat{f}(I) = 0$ by Lemma 2.1.

Lemma 4.2. *Let $f : \{0,1\}^n \to \{0,1\}$. Then there exist $c_1, c_2 > 0$ such that for all $I \subseteq [n]$ and arbitrary ϵ with $0 \le \epsilon \le 1$,*

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

where

$$\alpha_I = \tfrac{1}{2}\left(\mathrm{Var}(f) + \hat{f}(I)^2\right). \tag{3}$$

Proof. For $a, b \in \{0,1\}$, let $\alpha_I^{ab} = \Pr[x_I = a \wedge f(x) = b]$. It follows that $\alpha_I^{01} + \alpha_I^{11} = \Pr[f(x) = 1]$. We first prove that

$$\alpha_I = \alpha_I^{00}\alpha_I^{11} + \alpha_I^{10}\alpha_I^{01} \;. \tag{4}$$

Note that $\hat{f}(I) = 2^{-n}\sum_{x\in\{0,1\}^n} f(x)\cdot(-1)^{x_I} = \alpha_I^{01} - \alpha_I^{11}$. We have

$$\begin{aligned}
\alpha_I^{00}\alpha_I^{11} + \alpha_I^{10}\alpha_I^{01} - \tfrac{1}{2}Var(f) &= (\tfrac{1}{2} - \alpha_I^{01})\alpha_I^{11} + (\tfrac{1}{2} - \alpha_I^{11})\alpha_I^{01} - \tfrac{1}{2}\Pr[f(x)=0]\Pr[f(x)=1] \\
&= \tfrac{1}{2}\left(\alpha_I^{11} + \alpha_I^{01} - 4\alpha_I^{01}\alpha_I^{11} - \Pr[f(x)=1] + \Pr[f(x)=1]^2\right) \\
&= \tfrac{1}{2}\Pr[f(x)=1]^2 - 2\alpha_I^{01}\alpha_I^{11} \;=\; \tfrac{1}{2}(\alpha_I^{01} - \alpha_I^{11})^2 \;=\; \tfrac{1}{2}\hat{f}(I)^2 \;\;.
\end{aligned}$$

Let S be a uniformly distributed sample of size m and for $a, b \in \{0,1\}$, let A_I^{ab} denote the set of example indices k such that $(x_I^k, y^k) = (a, b)$. Since E_I has an edge between all pairs $(x^k, f(x^k))$ and $(x^\ell, f(x^\ell))$ with $x_I^k \neq x_I^\ell$ and $f(x^k) \neq f(x^\ell)$, 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 Chernoff bound (Fact 2.1), with probability at least $1 - 2e^{-2\delta m^2}$, $|A_I^{ab} - \alpha_I^{ab} m| \le \delta m$. Thus we can find $c_1, c_2 > 0$ such that

$$\left||E_I| - \alpha_I m^2\right| = \left||A_I^{00}|\cdot|A_I^{11}| + |A_I^{10}|\cdot|A_I^{01}| - (\alpha_I^{00}\alpha_I^{11} + \alpha_I^{10}\alpha_I^{01})m^2\right| < c_1 e^{-c_2\epsilon^2 m}$$

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

We should mention that in fact, for all $I \subseteq [m]$, the expectation of $|E_I|$ is equal to

$$(\alpha_I^{00}\alpha_I^{11} + \alpha_I^{10}\alpha_I^{01})m(m-1) = \tfrac{1}{2}\left(\mathrm{Var}(f) + \hat{f}(I)^2\right)m(m-1) \approx \alpha_I m^2 \;.$$

5 Analysis of Greedy

5.1 Greedy Succeeds for all Functions that are Fourier-Accessible

In this section, we state and prove our main results. Let us start with the positive result, the class of functions for which Greedy is successful.

Theorem 5.1. *There is a polynomial p such that for every Fourier-accessible function $f : \{0,1\}^n \to \{0,1\}$ and every $\delta > 0$, given a uniformly distributed sample S for f of size $m \ge p(2^d, \log n, \log(1/\delta))$, where $d = |\,\mathrm{rel}(f)|$,* Greedy *outputs exactly the variables in* $\mathrm{rel}(f)$ *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 2.3, with probability at least $1 - \delta/2$, *any* d-junta that is consistent with a sample S for f of size $m \ge m_0 = 2^{2d+1}\ln\frac{2n}{\delta}$ 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 $x_{i_1}, \ldots, x_{i_d}$. For $s \in [d]$, let $R_s = \{i_1, \ldots, i_s\}$. We prove that with high probability, each variable that is output is relevant to f. This implies that GREEDY halts exactly after d steps since E can always be covered by the sets E_i, $x_i \in \mathrm{rel}(f)$.

Let $\epsilon = 2^{-3d-3}$. For each $I \subseteq [n]$ with $1 \leq |I| \leq d$, we have $\Pr[||E_I| - \alpha_I m^2| > \epsilon m^2] < c_1 e^{-c_2 \epsilon^2 m}$ by Lemma 4.2. Consequently,

$$\forall I \subseteq [n] \text{ such that } 1 \leq |I| \leq d : \ ||E_I| - \alpha_I m^2| \leq \epsilon m^2 \tag{5}$$

with probability at least $\eta = 1 - n^d \cdot c_1 e^{-c_2 \epsilon^2 m}$. In the following, we assume that (5) holds. Thus, all subsequent consequences of (5) hold with probability at least η.

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 step of GREEDY, i.e., $E_i^{(s)} = E_i \setminus \{E_{i_1} \cup \ldots \cup E_{i_s}\}$.

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 step s, there is an $i^* \in \mathrm{rel}(f) \setminus R_s$ and an $I^* \subseteq R_s$ such that $\hat{f}(I^* \cup \{i^*\}) \neq 0$ and hence $|\hat{f}(I^* \cup \{i^*\})| \geq 2^{-d}$. Otherwise, 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 4.1 implies

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

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

It suffices to choose $m \geq m_1 = c_2^{-1} \cdot 2^{6d-6}(d \ln n + \ln(2c_1/\delta))$ to have $\eta \geq 1 - \delta/2$. 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. □

5.2 Greedy Fails for all Functions that are *not* Fourier-Accessible

Lemma 5.1. *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 $E_i^{(s)} = E_i \setminus (E_{i_1} \cup \ldots \cup E_{i_s})$ for $i \in [n]$. Given that $\{x_{i_1}, \ldots, x_{i_s}\} \subseteq \mathrm{acc}(f) \cup \mathrm{irrel}(f)$, 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 $|E_j^{(s)}|$, $x_j \notin \mathrm{rel}(f)$, 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) We show that for $a, b \in \{0,1\}$ and $v \in \{0,1\}^s$,

$$\Pr[x_i = a \mid (x_{i_1}, \ldots, x_{i_s}) = v \wedge f(x) = b] = \tfrac{1}{2} ,$$

just as for the irrelevant variables: Let $R = \{i_1, \ldots, i_s\}$ and $I \subseteq R$. If $I \subseteq \operatorname{acc}(f)$, then since x_i is not accessible, $\hat{f}(I \cup \{i\}) = 0$. Otherwise, I contains some irrelevant variable index, and hence $\hat{f}(I \cup \{i\}) = 0$ by Lemma 2.1. By Lemma 2.2, $\widehat{f_{R \leftarrow v}}(i) = 0$ for all $v \in \{0,1\}^s$. Since

$$\begin{aligned} \widehat{f_{R \leftarrow v}}(i) &= \Pr[x_i = 0 \wedge f_{R \leftarrow v}(x) = 1] - \Pr[x_i = 1 \wedge f_{R \leftarrow v}(x) = 1] \\ &= \Pr[x_i = 1 \wedge f_{R \leftarrow v}(x) = 0] - \Pr[x_i = 0 \wedge f_{R \leftarrow v}(x) = 0] \ , \end{aligned}$$

and since $\Pr[x_i = 0 \wedge f_{R \leftarrow v}(x) = b] + \Pr[x_i = 1 \wedge f_{R \leftarrow v}(x) = b] = \Pr[f_{R \leftarrow v}(x) = b]$, we have $\Pr[x_i = a \wedge f_{R \leftarrow v}(x) = b] = \frac{1}{2}\Pr[f_{R \leftarrow v}(x) = b]$. Consequently, writing x^R for $(x_{i_1}, \ldots, x_{i_s})$, we obtain

$$\begin{aligned} \Pr[x_i = a \wedge x^R = v \wedge f(x) = b] &= \Pr[x_i = a \wedge f(x) = b \mid x^R = v] \cdot \Pr[x^R = v] \\ &= \Pr[x_i = a \wedge f_{R \leftarrow v}(x) = b] \cdot \Pr[x^R = v] \\ &= \frac{1}{2}\Pr[f_{R \leftarrow v}(x) = b] \cdot \Pr[x^R = v] \\ &= \tfrac{1}{2}\Pr[f(x) = b \mid x^R = v] \cdot \Pr[x^R = v] \\ &= \frac{1}{2}\Pr[f(x) = b \wedge x^R = v] \ . \end{aligned}$$

Thus, $\Pr[x_i = a \mid x^R = v \wedge f(x) = b] = 1/2$, 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 $|E_i^{(s)}|$ and all $|E_j^{(s)}|$, $x_j \notin \operatorname{rel}(f)$, 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

$$x_{i_{s+1}} \in \{x_i\} \cup (\operatorname{irrel}(f) \setminus \{x_{i_1}, \ldots, x_{i_s}\}) .$$

Since all cardinalities $|E_i^{(s)}|$ corresponding to the variables in $\{x_i\} \cup (\operatorname{irrel}(f) \setminus \{x_{i_1}, \ldots, x_{i_s}\})$ are identically distributed, the probability that x_i is selected then is at most $1/(|\operatorname{irrel}(f) \setminus \{x_{i_1}, \ldots, x_{i_s}\}| + 1)$. Hence the probability that $x_{i_{s+1}}$ is relevant but not accessible is at most

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

□

The following negative result, the class of functions for which GREEDY fails, complements Theorem 5.1.

Theorem 5.2. *Let $f : \{0,1\}^n \to \{0,1\}$ be a concept that is not Fourier-accessible and $\lambda \geq 1$. Given a sample S for f of arbitrary size,* GREEDY *λ-fails on input S with probability at least $1 - \frac{\lambda d^2}{n - \lambda d}$, where $d = |\operatorname{rel}(f)|$.*

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 $\{x_{i_1}, \ldots, x_{i_s}\} \subseteq \mathrm{acc}(f) \cup \mathrm{irrel}(f)$ *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* $\{x_{i_1}, \ldots, x_{i_s}\} \subseteq \mathrm{acc}(f) \cup \mathrm{irrel}(f)$, which is at most $d/(|\,\mathrm{irrel}(f) \setminus \{x_{i_1}, \ldots, x_{i_s}\}|)$ by Lemma 5.1 (b).

Suppose that GREEDY is λ-successful on input S, i.e., $t \leq \lambda \cdot d$ and $\{x_{i_1}, \ldots, x_{i_t}\} \supseteq \mathrm{rel}(f)$. Since f is not Fourier-accessible, there exists $s \in \{0, \ldots, t-1\}$ such that $x_{i_1}, \ldots, x_{i_s} \in \mathrm{acc}(f) \cup \mathrm{irrel}(f)$ 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 $1 - \frac{\lambda d^2}{n - \lambda d}$. □

Note that Theorem 5.2 not only says that GREEDY (with high probability) fails for concepts that are not Fourier-accessible, but that GREEDY even fails to find all relevant variables of the target concept in $\lambda \cdot |\,\mathrm{rel}(f)|$ rounds for any $\lambda \geq 1$. In addition, note that the claim in Theorem 5.2 is independent of the sample size.

In the past, it has often been emphasized that GREEDY has a "logarithmic approximation guarantee" (see [1, 3, 11, 16]), i.e., given a sample S for f of size m, GREEDY finds a set of at most $(2 \ln m\ + 1) \cdot |\,\mathrm{rel}(f)|$ variables that explain S. Theorem 5.2 shows that if f is not Fourier-accessible, then with probability at least $\frac{(2 \ln m + 1)d^2}{n - (2 \ln m + 1)d}$, these variables *do not contain all relevant variables* (where $d = |\,\mathrm{rel}(f)|$). Thus, given a sample that originates from 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.

6 Concluding Remarks

6.1 Variations of the Greedy Algorithm

In previous work [6], we have introduced the *gap* $\Delta_i(f, D)$ of variable x_i with respect to the target concept f and the attribute distribution D. It has been shown that the GREEDY RANKING algorithm (see Section 1) is successful provided that $\Delta_i(f, D) > 0$ for all $x_i \in \mathrm{rel}(f)$. If D is the uniform distribution, one can prove using (4) that $\Delta_i(f, D) = \alpha_i - \frac{1}{2}\mathrm{Var}(f) = \frac{1}{2}\hat{f}(i)^2$. Hence $\Delta_i(f, D) > 0$ if and only if $\hat{f}(i) \neq 0$. Moreover, $\Delta_i(f, D) > 0$ for all $x_i \in \mathrm{rel}(f)$ iff f is 1-low. Similar reasoning to the proof of Theorem 5.2 shows that for functions that are *not* 1-low, GREEDY RANKING fails with high probability. Clearly, all 1-low functions are Fourier-accessible. A simple example of a function that is Fourier-accessible but not 1-low is given by the function f_1 in Table 1. Thus, Fourier-accessibility is strictly weaker than 1-lowness – and hence by Theorem 5.1, GREEDY can cope with a strictly larger class than what has been provided by Fukagawa and Akutsu [16].

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 $\mathrm{NAE} : \{0,1\}^d \to \{0,1\}$ defined by $\mathrm{NAE}(x) = 1$ iff there exist $i, j \in [d]$ such that $x_i \neq x_j$. 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 t-low juntas for $t > 1$. Allowing GREEDY RANKING to also select new attributes x_I with $|I| \leq t$, yields an algorithm that is applicable exactly to the class of t-low juntas. Thus, such an algorithm provides an analog to the Fourier-based algorithm that simply checks for each Fourier coefficient $\hat{f}(I)$, $I \subseteq [n]$ with $1 \leq |I| \leq t$, whether it vanishes or not, see [22, 7].

We conjecture that allowing GREEDY to use attributes x_I with $|I| \leq t$ yields an algorithm that can cope exactly with the class of *t-Fourier-accessible* functions, where t-Fourier-accessible

is defined in the same way as Fourier-accessible except that in item 2 of Definition 2.2, we allow $|I_j \setminus I_{j-1}| \leq t$ for all $j \in [s]$. While the generalization of Theorem 5.1 is straightforward, finding an appropriate analog of Theorem 5.2 seems to need more careful reasoning.

6.2 Multi-Valued Attributes and Non-Uniform Distribution

Since it is straightforward to extend the greedy algorithms to multi-valued attributes, it is natural to ask whether there is still a connection between the cardinalities $|E_i|$ and the corresponding Fourier coefficients. Indeed, using Fourier analysis of functions on finite Abelian groups, we can prove that for $f : \mathbb{Z}_r^n \to \{0,1\}$, if $\Delta_i(f, D) = 0$, then also $\hat{f}(e_i) = 0$. Although the converse is false in general, it does hold if r is prime.

Quantitatively, we can show that $\Delta_i(f) \in \Theta(|\hat{f}(i)|^2)$ for $r \in \{2, 3, 4, 6\}$, whereas somewhat oddly, for $r = 5$ or $r \geq 7$, there are functions such that $|\hat{f}(e_i)|^2 \in o(\Delta_i(f))$. The latter circumstances are essentially due to the fact that for $r \in \{2, 3, 4, 6\}$, $\mathbb{Z}[\omega_r]$ is a discrete lattice in $\mathbb{C}$, whereas for $r = 5$ or $r \geq 7$, $\mathbb{Z}[\omega_r]$ is dense in $\mathbb{C}$.

For non-uniform attribute distributions – although a generic notion of Fourier coefficients can be given [8, 17] – Lemma 4.2 with a similar definition of α_I does not hold any more. In fact, it is easy to find examples such that

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

(b) there are $x_i \in \mathrm{rel}(f)$ and $x_j \in \mathrm{irrel}(f)$ such that the expected sizes of E_i and E_j are equal, although $\hat{f}(i) = \hat{f}(j) = 0$.

6.3 Further Open Problems

The first issue left for future research is the investigation of the performance of the greedy algorithm in variations of the learning scenario considered in this paper: attributes and classifications may take more than two values, attributes may be non-uniformly distributed, the given data may contain noise, etc.

The second issue is to stick to the learning scenario 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 [4] 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 $n \gg |\mathrm{rel}(f)|$, 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.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] Noga Alon and Joel Spencer. *The Probabilistic Method*. Wiley-Intersci. Ser. Discrete Math. Optim. John Wiley and Sons, 1992.

[6] 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.

[7] 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 on Computation and Logic, TAMC06, Beijing, May, 2006*, volume 3959 of *Lecture Notes in Comput. Sci.*, pages 387–398, 2006.

[8] R. R. 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.

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

[10] Avrim Blum, Merrick Furst, Jeffrey 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.

[11] Avrim Blum and Pat Langley. Selection of Relevant Features and Examples in Machine Learning. *Artificial Intelligence*, 97(1-2):245–271, 1997.

[12] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Occam's razor. *Inform. Process. Lett.*, 24(6):377–380, 1987.

[13] 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.

[14] Vasek Chvatal. A Greedy Heuristic for the Set Covering Problem. *Math. Oper. Res.*, 4(3):233–235, 1979.

[15] Uriel Feige. A threshold of $\ln n$ for approximating set cover. *J. ACM*, 45(4):634–652, 1998.

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

[17] 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.

[18] David Johnson. Approximation Algorithms for Combinatorial Problems. *J. Comput. System Sci.*, 9(3):256–278, 1974.

[19] Jon Kleinberg and Éva Tardos. *Algorithm Design*. Addison Wesley, 2005.

[20] Nathan Linial, Yishay Mansour, and Noam Nisan. Constant Depth Circuits, Fourier Transform, and Learnability. *J. ACM*, 40(3):607–620, 1993.

[21] 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.

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

[23] 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.

[24] 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.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de/