Learning DNF through Generalized Fourier Representations

Mohsen Heidari* MHEIDAR@IU.EDU and **Roni Khardon*** RKHARDON@IU.EDU
Department of Computer Sciences, Indiana University, Bloomington, IN, USA

Editors: Nika Haghtalab and Ankur Moitra

Abstract

The Fourier representation for the uniform distribution over the Boolean cube has found numerous applications in algorithms and complexity analysis. Notably, in learning theory, the learnability of Disjunctive Normal Form (DNF) under the uniform and product distributions has been established through such representations. This paper makes three main contributions. First, it introduces a generalized Fourier expansion that can be used with any distribution D through the representation of the distribution as a Bayesian network (BN). Second, it shows that the main algorithmic tools for learning with the Fourier representation that use membership queries to approximate functions by recovering their heavy Fourier coefficients, can be used with slight modifications with the generalized expansion. These results hold for any distribution. Third, it analyzes the L_1 spectral norm of conjunctions under the new expansion, showing that it is bounded for a class of distributions which can be represented by a difference-bounded tree BN, where a parent node in the BN representation can change the conditional expectation of a child node by at most $\alpha < 0.5$. Lower bounds are presented to show that such constraints are necessary. Combining these contributions, the paper shows learnability of DNF with membership queries under difference-bounded tree BN.

Keywords: PAC Learning; Membership Queries; Fourier Basis; DNF

1. Introduction

The problem of learning Disjunctive Normal Form (DNF) expressions from examples has been a major open problem since its introduction by Valiant (1984). Significant progress has been made by considering subclasses of expressions (e.g., (Valiant, 1985; Bshouty and Tamon, 1996; Sakai and Maruoka, 2000)) or making distribution assumptions (Verbeurgt, 1990; Servedio, 2004), and the best-known algorithm for the general problem is not polynomial time (Klivans and Servedio, 2004). A potentially less demanding model allows for an additional source of information through membership queries (MQ). In this model, in addition to random examples, the learner can ask for the label of the target function on any input. Valiant (1984) gave an efficient MQ learning algorithm for Monotone DNF. Since then, several subclasses of DNF have been shown to be learnable in this model (e.g. (Bshouty, 1995; Kushilevitz, 1996; Aizenstein et al., 1998; Hellerstein et al., 2012)).

Angluin and Kharitonov (1995) have shown that (under cryptographic assumptions) the general distribution free case for general DNF is not easier with MQ. On the other hand, positive results have been obtained for specific distributions. Jackson (1997) gave the first polynomial time MQ learning algorithm for DNF over c-bounded product distributions. This result was based on the Fourier representation of functions over the Boolean cube (Linial et al., 1993) and combines the algorithm by Kushilevitz and Mansour (henceforth KM algorithm) for finding the heavy Fourier coefficient of a boolean function (Goldreich and Levin, 1989; Kushilevitz and Mansour, 1993) with Boosting. The approach was elaborated and improved by several authors (Bshouty et al., 2004;

* Equal Contribution.

Feldman, 2007; Kalai et al., 2009; Feldman, 2012). However, to date, the Fourier approach has been largely limited to product distributions and implications for DNF learnability are restricted to such distributions. In this paper we provide a significant extension of these results to a broad class of distributions. To achieve this the paper makes several distinct contributions.

First, we develop a novel generalized Fourier representation induced by any distribution D, *by using the Bayesian Network (BN) representation of* D. A BN specifies a distribution using a directed acyclic graph (DAG) and conditional probability distributions where each node is conditioned on its parents (Pearl, 1989; Koller and Friedman, 2009). The generalized Fourier expansion constructs basis functions ϕ_S for $S \subseteq \{1, \ldots, n\}$ using the graph structure and conditional probabilities that specify the BN, yielding an orthonormal basis so that for any function we have $f(\mathbf{x}) = \sum_S \hat{f}_S \phi_S(\mathbf{x})$. While the new basis preserves some important properties, unlike the standard construction, it does not impose sparsity. That is, if f depends only on a subset of variables T and $S \setminus T \neq \emptyset$, the value of the coefficient $\hat{f}_S$ may not be zero.

The second contribution is showing that the KM algorithm can be extended for the new basis and with high probability it recovers all the large coefficients, $|\hat{f}_S| \geq \theta$, of a function f. The algorithm does not require inference with the BN (e.g. calculating marginal or conditional probabilities) that can be computationally hard, but only requires forward sampling which is always feasible.

Our third contribution is in analyzing the Fourier representation of conjunctions g with d literals, specifically providing bounds for its spectral norm $L_1(g) = \sum_S |\hat{g}_S|$. This type of analysis is easy for the uniform case (where $L_1 = 1$ (Blum et al., 1994; Khardon, 1994)) and the product case (where $L_1 = O(2^{d/2})$ (Feldman, 2012)), but is nontrivial for general distributions due to the fact that sparsity does not hold. We first analyze the spectral norm of conjunctions under k-junta distributions (Aliakbarpour et al., 2016) showing that $L_1(g) = O(2^{(k+d)/2})$. Then, turning to general distributions, we show that the values of the coefficients are determined in a combinatorial manner by the values of the corresponding BN parameters. We then derive bounds for a broad class of difference bounded tree BN distributions, where the value of a parent can change the conditional probability of a child by at most $\alpha < 0.5$ (noting that product distributions satisfy this with $\alpha = 0$). In particular, in chain BNs we have $L_1(g) = O((\frac{2}{1-2\alpha})^d)$ and for tree BN we have $L_1(g) = O((\frac{2}{1-2\alpha})^{2d})$. The upper bounds are complemented by showing that without boundedness or without a tree structure the spectral norm can be exponentially large even for $d = 1$. As a byproduct, our analysis also provides an exact value for the spectral norm in the product case.

With these in place, we obtain our main results, showing that the extended KM algorithm can be directly used to learn decision trees and the algorithm PTFconstruct of Feldman (2012) can be used with a slight modification (removing the filtering of large degree coefficients, which is only suitable with sparsity) to learn DNF, under the corresponding families of distributions.

Summary of the contributions: To summarize, the paper develops a generalized Fourier basis and shows that major algorithmic tools from learning theory can be used with this basis. Using these and an analysis of the spectral norm for conjunctions, the paper shows the learnability of DNF under difference-bounded distributions, significantly extending previous results to a broad class of distributions. We emphasize that the basis and the extended KM algorithm are valid for any distribution and they do not require a tree structure or boundedness. These conditions are required only to establish spectral norm bounds used for learnability results. More specifically, our main contributions include:

- A new Fourier basis for any distribution D by using the BN representation of D (Definition 3).

- An extension of the KM algorithm to any distribution using the new basis (Theorem 7), a proof of learnability of disjoint DNFs using KM algorithm (Corollary 24) and learnability of DNFs using the PTFconstruct algorithm of Feldman (2012) (Corollary 25).

- An exact Fourier expansion of conjunctions under chain BNs (Lemma 15) and an upper bound on the spectral norm of d literal conjunctions for difference-bounded chains (Theorem 17).

- An exact characterization of the spectral norm of d literal conjunctions under product distributions, and a tighter upper bound on the spectral norm (Proposition 18).

- An upper bound for the spectral norm of d literal conjunctions under difference bounded tree BNs (Theorem 19) and k-junta distributions (Lemma 8).

- An exponential lower bounds on the spectral norm of 1-literal conjunctions, when the difference-bounded condition is violated, or without the tree BN structure (Lemma 22 and Theorem 23).

Due to space constraints, some of the constructions or proof details are omitted here; these are provided in the full version of the paper (Heidari and Khardon, 2025).

Finally, note that the discussion so far assumed that a BN representation of the distribution is given to the learner. While learning general distributions is computationally hard, it is well known that tree BN are learnable (Chow and Liu, 1968; Höffgen, 1993; Bhattacharyya et al., 2023). The full version of the paper shows that the learning algorithm can be modified to learn difference bounded distributions, and combined with the results above to enable learnability of DNF even when the distribution is not known in advance.

Organization of the paper: We start with basic definitions and the preliminaries in Section 2. The new BN induced Fourier basis is introduced in Section 3 and the extension of the KM algorithm to arbitrary distributions is presented in Section 4. The Fourier expansion of conjunctions and upper bounds on the spectral norm under k-junta distributions, chain BNs, and tree BNs, are discussed in sections 5, 6, and 7, respectively, and Section 8 presents the lower bounds on the spectral norm. The implications of our results for the learnability of DNFs are discussed in Section 9. Lastly, Section 10 concludes with a summary and some questions for future work.

2. Preliminaries

Notations. We use $[n]$ to denote the sequence $[1, 2, \ldots, n]$, $[a, b)$ to denote the sequence $a, a+1, \ldots, b-1$, and (a, b) and $[a, b)$ are defined similarly. Capital letters are used for random variables and lower case letters to denote their assignments.

Learning in this paper is defined based on the well-known *probably approximately correct* (PAC) model (Valiant, 1984; Kearns et al., 1994). In this framework, the learner finds an approximation h to an unknown target Boolean function f, given an oracle access to f in the form of labeled examples $(\mathbf{x}, f(\mathbf{x}))$ with $\mathbf{x}$ generated based on an unknown distribution D. With membership queries (MQ), in addition to random examples, the algorithm can ask for the label $f(\mathbf{x})$ of any example $\mathbf{x}$ of its choice. The objective is to output a hypothesis h that is close to the target function f in terms of its predictions. More formally, class $\mathcal{F}$ is learnable with MQ if $\forall f \in \mathcal{F}$, with oracle and MQ access to f, with high probability the algorithm outputs h such that $P_{\mathbf{x}\sim D}(h(\mathbf{x}) \neq f(\mathbf{x})) \leq \epsilon$.

Bayesian Network. The key to our analysis is incorporating the representation of the distribution D as a BN into the representation of functions. Since BN is a universal representation, this does not restrict the set of distributions under consideration. A BN is specified by a directed acyclic graph and a set of conditional probability tables (or functions). Let $G = (V, E)$ be the graph where nodes correspond to the individual random variables. Then, the joint probability distribution can be written as a product of individual probability distributions of each node conditioned on its parent variables: $P(X_1, \cdots, X_n) = \prod_{v \in V} P(X_v | X_{\mathrm{pa}(v)})$, where $\mathrm{pa}(v)$ is the set of the parents of node v and where in this paper the variables are binary, i.e., $X_i \in \{0, 1\}$.

For some of the results, we will need to restrict the class of distributions D. For any node v, let $\mu_{v,x_{\mathrm{pa}(v)}}$ and $\sigma_{v,x_{\mathrm{pa}(v)}}$ denote the conditional expectation and standard deviation of X_v given its parents' realization $x_{\mathrm{pa}(v)}$, respectively. Note that the values $\{\mu_{v,x_{\mathrm{pa}(v)}}\}$ are exactly the parameters that specify the BN representation of D.

Definition 1 *A distribution D is* $c-$bounded *for $c \in (0, 1)$ if for all $v \in V$ and any assignments x, x' we have $c \leq P(X_v = x | x_{\mathrm{pa}(v)} = x') \leq 1 - c$. Moreover, D is* α-difference-bounded *if for all v and for any two assignments $x_{\mathrm{pa}(v)}, y_{\mathrm{pa}(v)}$ to $\mathrm{pa}(v)$, we have $\left|\mu_{v,x_{\mathrm{pa}(v)}} - \mu_{v,y_{\mathrm{pa}(v)}}\right| \leq \alpha$ and $\left|\sigma_{v,x_{\mathrm{pa}(v)}} - \sigma_{v,y_{\mathrm{pa}(v)}}\right| \leq \alpha$.*

Based on this definition, a distribution is a *difference-bounded tree BN*, if it can be expressed as a BN where each node has at most a single parent and the conditional probability tables in the specification are c-bounded and α-difference bounded. For some of the results we need the following observation that follows trivially for c-bounded distributions:

Lemma 2 *For any c-bounded distribution D and for any conjunction f with d literals, $c^d \leq \mathbb{E}_D[f(\mathbf{X})] \leq (1 - c)^d$.*

Fourier Expansion on the Boolean Cube with Product Distributions. Our results generalize previous works that established learnability of functions using the Fourier representation under uniform and bounded product distributions (Linial et al., 1993; Kushilevitz and Mansour, 1993; Furst et al., 1991; Feldman, 2012). These distributions are captured by BNs with an empty edge set. The boundedness condition holds and the differences are zero and hold trivially.

We briefly discuss these constructions; see (O'Donnell, 2014; Wolf, 2008) for a review. In the following, given any distribution D, define the induced inner product between any pair of functions $f, g : \{0, 1\}^n \to \mathbb{R}$ by $\langle f, g \rangle = \mathbb{E}_D[f(\mathbf{X})g(\mathbf{X})]$. For the uniform distribution, let $\psi_i(\mathbf{x}) = (-1)^{x_i} = 1 - 2x_i$ where $x_i \in \{0, 1\}$ and $i \in [n]$. The basis, defined as $\psi_S(\mathbf{x}) = \prod_{i \in S} \psi_{x_i}(\mathbf{x})$, for all subsets $S \subseteq [n]$ and $\mathbf{x} \in \{0, 1\}^n$ is orthonormal under the uniform distribution. As a result, any function f on the Boolean cube admits the decomposition $f(\mathbf{x}) = \sum_{\mathcal{S} \subseteq [n]} \hat{f}_{\mathcal{S}}\, \psi_{\mathcal{S}}(\mathbf{x})$, where $\hat{f}_{\mathcal{S}}$ are called the Fourier coefficients of f and are calculated as

$$\hat{f}_{\mathcal{S}} = \langle f, \psi_{\mathcal{S}} \rangle = \mathbb{E}_{\mathbf{X} \sim \mathrm{Uniform}}[f(\mathbf{X})\psi_{\mathcal{S}}(\mathbf{X})] = \frac{1}{2^n} \sum_{\mathbf{x}} f(\mathbf{x})\psi_{\mathcal{S}}(\mathbf{x}).$$

The basis for product distributions, generalizing this, is given by $\psi_i(\mathbf{x}) = \frac{\mu_i - x_i}{\sigma_i}$, where μ_i and σ_i are the expectation and the standard deviation of X_i. Here too, the basis is orthonormal with $\hat{f}_{\mathcal{S}} = \mathbb{E}_D[f(\mathbf{X})\psi_{\mathcal{S}}(\mathbf{X})]$. This formulation reduces to the one for the uniform distribution, where $\mu_i = 0.5$ and $\sigma_i = \sqrt{\mu_i(1 - \mu_i)} = 0.5$.

The above expansion for non-uniform product distributions is not technically a Fourier transform as $\psi_{\mathcal{S}}$ are not group characters and therefore some properties are not satisfied (see (O'Donnell, 2014)). Therefore, we use the term *Fourier expansion* to distinguish it from the Fourier transform. We note that recent work studied Fourier expansion to non-product distributions (Heidari et al., 2021, 2022). However, since the basis in that work is learned from data, it is not clear how to apply it for learning of Boolean functions, and specifically to the analysis of conjunctions.

3. BN Induced Fourier Basis

In what follows, we define a distribution dependent Fourier basis for functions on the Boolean cube. The key is to define the basis using not just the distribution but using a specific representation of the distribution as a BN. For a BN G, with $|V| = n$, we can identify the nodes with their indices, i.e., $V = \{1, \ldots, n\}$. For any node $v \in V$ in a BN G, define $\phi_v(\mathbf{x}) := \frac{x_v - \mu_{v,x_{\mathrm{pa}(v)}}}{\sigma_{v,x_{\mathrm{pa}(v)}}}$. We then define:

Definition 3 (BN Induced Fourier Basis) *The Boolean Fourier basis for a BN given by G and the associated parameters is defined as follows: For all $\mathcal{S} \subseteq V$ and $\mathbf{x} \in \{0,1\}^n$.*

$$\phi_{\mathcal{S}}(\mathbf{x}) := \prod_{v \in \mathcal{S}} \phi_v(\mathbf{x}) = \prod_{v \in \mathcal{S}} \frac{x_v - \mu_{v,x_{\mathrm{pa}(v)}}}{\sigma_{v,x_{\mathrm{pa}(v)}}}. \tag{1}$$

To simplify the presentation, we assume G is known and omit it from the notation. The next lemma shows that this is indeed an orthonormal basis and develops some more useful properties:

Lemma 4 *The following holds for any $\mathcal{S}, \mathcal{T} \subseteq [n]$ and $\mathbf{x} \in \{0,1\}^n$.*
(a) $\mathbb{E}[\phi_{\mathcal{S}}(\mathbf{X})] = 0$, and $\mathbb{E}[\phi_{\mathcal{S}}(\mathbf{X})|\mathbf{x}_{\mathrm{pa}(\mathcal{S})}] = 0$, where $\mathrm{pa}(\mathcal{S})$ is the set of parents of all $v \in \mathcal{S}$.
(b) $\mathbb{E}[\phi^2_{\mathcal{S}}(\mathbf{X})] = 1$, and $\mathbb{E}[\phi^2_{\mathcal{S}}(\mathbf{X})|x_{\mathrm{pa}(\mathcal{S})}] = 1$.
(c) $\phi_{\mathcal{S}}\phi_{\mathcal{T}} = \phi^2_{\mathcal{S}\cap\mathcal{T}}\phi_{\mathcal{S}\Delta\mathcal{T}}$.
(d) $\mathbb{E}[\phi_{\mathcal{S}}(\mathbf{X})\phi_{\mathcal{T}}(\mathbf{X})] = 0$ and $\mathbb{E}[\phi_{\mathcal{S}}(\mathbf{X})\phi_{\mathcal{T}}(\mathbf{X})|\mathbf{x}_{\mathrm{pa}(\mathcal{S}\cup\mathcal{T})}] = 0$ for $\mathcal{T} \neq \mathcal{S}$.

Proof Sketch: It is easy to check that individual basis functions are normalized $\mathbb{E}_v\left[\phi_v(\mathbf{x})|X_{\mathrm{pa}(v)}\right] = (\mu_{v,X_{\mathrm{pa}(v)}} - \mu_{v,X_{\mathrm{pa}(v)}})/\sigma_{v,X_{\mathrm{pa}(v)}} = 0$ and similarly $\mathbb{E}_v\left[\phi_v(\mathbf{x})^2|X_{\mathrm{pa}(v)}\right] = 1$. For $\mathcal{S} = \{j_1, \ldots, j_k\}$ where $j_1, j_2, \ldots, j_k$ satisfy the BN ordering,

$$\mathbb{E}[\phi_{\mathcal{S}}(\mathbf{X})|x_{\mathrm{pa}(\mathcal{S})}] = \mathbb{E}_{X_{j_1}}[\phi_{j_1}(\mathbf{X})\,\mathbb{E}_{X_{j_2}}[\phi_{j_2}(\mathbf{X})\,\ldots\mathbb{E}_{X_{j_k}}[\phi_{j_k}(\mathbf{X})|x_{\mathrm{pa}(X_{j_k})}]\ldots|x_{\mathrm{pa}(X_{j_2})}]|x_{\mathrm{pa}(X_{j_1})}].$$

Performing the expectation in reverse order from k to 1 yields 0, and similarly for ϕ^2 it yields 1. This proves (a,b). Property (c) holds by definition and using it we can prove (d) in the same manner as (b). ■

Properties (b,d) show that the functions $\phi_{\mathcal{S}}$ form an orthonormal basis w.r.t. inner product $\langle f, g\rangle = \mathbb{E}_D[f(\mathbf{X})g(\mathbf{X})]$. Therefore, for all functions $f : \{0,1\}^n \to \mathbb{R}$, $f(\mathbf{x}) = \sum_{\mathcal{S}} \hat{f}_{\mathcal{S}}\phi_{\mathcal{S}}(\mathbf{x})$, where $\hat{f}_{\mathcal{S}} = \mathbb{E}_D[f(\mathbf{X})\phi_{\mathcal{S}}(\mathbf{X})]$ is the Fourier coefficient of f.

The following definition is instrumental for the analysis of learnability:

Definition 5 *The spectral norm of a function f under a distribution D is the sum of the absolute values of its Fourier coefficients under D, $L_1(f) := \sum_{\mathcal{S}} |\hat{f}_{\mathcal{S}}|$.*

4. Extending KM to Arbitrary Distributions

In this section, we show that the KM algorithm (Kushilevitz and Mansour, 1993) can be generalized to recover the large coefficients with the new basis. The main new development in our work is given in the following lemma, which shows how the construction of the main tool and its proof of correctness can be adapted to the new basis.

To introduce the generalization, we extend the notation to specify sets using binary strings. We represent a set $\mathcal{S} \subseteq [n]$ with a binary vector $\gamma \in \{0,1\}^n$ where $\gamma_i = 1$ if $i \in \mathcal{S}$, and $\gamma_i = 0$ otherwise. With this notation strings can refer to sets in a 1-1 manner in a natural way.

Let $\alpha \in \{0,1\}^k$ and $\beta \in \{0,1\}^{n-k}$. We use $\beta\alpha$ to denote the concatenation of the two strings which is of length n. To avoid confusion, our strings, serving as subscripts or arguments to our function f, basis functions or coefficients, will always be of length n. To achieve this we pad a string $\alpha \in \{0,1\}^k$ on the left with $\bar{0} = 0^{n-k}$ to get $\overline{0\alpha} = \bar{0}\alpha$. Similarly, we pad $\beta \in \{0,1\}^{n-k}$ on the right with $\bar{0} = 0^k$ to get $\overline{\beta 0} = \beta\bar{0}$. Let $\alpha \in \{0,1\}^k$ and define the function

$$g_\alpha(u) = \sum_{\beta \in \{0,1\}^{n-k}} \hat{f}_{\beta\alpha}\, \phi_{\overline{\beta 0}}(\overline{u0}), \tag{2}$$

for every $u \in \{0,1\}^{n-k}$. The next lemma which is a generalization of Lemma 3.2 of (Kushilevitz and Mansour, 1993) shows that $g_\alpha(u)$ can be computed as an expectation:

Lemma 6 *For any $\alpha \in \{0,1\}^k$ and $u \in \{0,1\}^{n-k}$, let $\mathbf{Y} = (X_{n-k+1}, \ldots, X_n)$ be the last k variables of $\mathbf{X}^n$, then $g_\alpha(u) = \mathbb{E}_{(\mathbf{Y}|(X_1,\ldots,X_{n-k})=u)}[f(u\mathbf{Y})\phi_{\overline{0\alpha}}(u\mathbf{Y})]$.*

Proof Starting from the RHS, by the Fourier expansion of $f(u\mathbf{Y})$ we have:

$$\mathbb{E}_{\mathbf{Y}|u}[f(u\mathbf{Y})\phi_{\overline{0\alpha}}(u\mathbf{Y})] = \mathbb{E}_{\mathbf{Y}|u}[\sum_{a_1}\sum_{a_2} \hat{f}_{a_1a_2}\phi_{a_1a_2}(u\mathbf{Y})\phi_{\overline{0\alpha}}(u\mathbf{Y})],$$

where $a_1 \in \{0,1\}^{n-k}$ and $a_2 \in \{0,1\}^k$. Based on Lemma 4, $\phi_{a_1a_2}\phi_{\overline{0\alpha}} = \phi_{\overline{a_1 0}}(\phi_{\bar{0}(a_2\wedge\alpha)})^2\phi_{\bar{0}(a_2\oplus\alpha)}$, where $\wedge, \oplus$ denote the AND and XOR operations on the binary vectors, respectively. Therefore,

$$RHS = \sum_{a_1}\sum_{a_2} \hat{f}_{a_1a_2}\phi_{\overline{a_1 0}}(\overline{u0})\mathbb{E}_{\mathbf{Y}|u}[(\phi_{\bar{0}(a_2\wedge\alpha)}(u\mathbf{Y}))^2\phi_{\bar{0}(a_2\oplus\alpha)}(u\mathbf{Y})],$$

where the last equality holds because active indices in $\overline{a_1 0}$ (as well as their ancestors) are restricted to the first $n-k$ variables. Therefore $\phi_{\overline{a_1 0}}(u\mathbf{Y}) = \phi_{\overline{a_1 0}}(\overline{u0})$ and we can pull this term out of the expectation. We can now proceed in evaluating the expectation over the active variables in $\bar{0}(a_2 \wedge \alpha), \bar{0}(a_2 \oplus \alpha)$ in reverse lexicographical order over the BN (i.e., from children to parents). In that sequential computation, for indices in $\bar{0}(a_2 \wedge \alpha)$, $\mathbb{E}_{X_j|X_{\mathrm{pa}(j)}}[\phi_j(u\mathbf{Y})^2|x_{\mathrm{pa}(j)}] = 1$ and for indices in $\bar{0}(a_2 \oplus \alpha)$, $\mathbb{E}_{X_j|X_{\mathrm{pa}(j)}}[\phi_j(u\mathbf{Y})|x_{\mathrm{pa}(j)}] = 0$. That is, the expectation is zero when $a_2 \neq \alpha$ and it is 1 when $a_2 = \alpha$. Therefore, as claimed, $RHS = \sum_{a_1} \hat{f}_{a_1\alpha}\phi_{\overline{a_1 0}}(\overline{u0}) = g_\alpha(u)$. ■

The above lemma is crucial in extending prior results (Kushilevitz and Mansour, 1993; Bellare, 1991; Jackson, 1997) to general BN distributions. The remaining details follow with small modifications. In particular, the algorithm requires an estimate of $\mathbb{E}_U[g_\alpha^2(U)]$. According to the lemma, this can be done with sampling through $\mathbb{E}_U[g_\alpha^2(U)] = \mathbb{E}_{U,Y_1|U,Y_2|U}[f(UY_1)\phi_{\overline{0\alpha}}(UY_1)f(UY_2)\phi_{\overline{0\alpha}}(UY_2)]$.

What is crucial for our case is that this only requires forward sampling in the BN: u is sampled sequentially from the roots of the BN and y_1, y_2 are sampled conditional on u, and they are conditionally independent. That is, the process can be performed in so called ancestral sampling (Koller and Friedman, 2009) which can be done in polynomial time. In addition, note from the lemma that α must be constructed backward (starting with leaves in BN order).

One additional complication arises for the estimation process. Denoting $Z = f(UY_1)\phi_{\overline{0\alpha}}(UY_1) f(UY_2)\phi_{\overline{0\alpha}}(UY_2)$ our goal is to estimate $\mathbb{E}[Z]$ but the range of Z can be exponential in n so we cannot use Hoeffding bounds as in prior work. Instead we directly estimate $G_\alpha = \frac{1}{m}\sum Z^i$ where Z^i are independent samples of Z. By first analyzing $\mathsf{var}(Z|u)$ which is ≤ 1 we can show that $\mathsf{var}(Z) \leq 5/4$. We can then use Chebyshev's bound to show that with high probability $|G_\alpha - \mathbb{E}_U[g_\alpha^2(U)]| \leq \theta^2/4$. Hence the algorithm is the same as the original KM algorithm, but it specifically constrains the variable order in the construction of α and the sampling process in estimating $\mathbb{E}_U[g_\alpha^2(U)]$. The analysis of remaining details is identical to previous work and yields:

Theorem 7 (cf. Theorem 3.10 of (Kushilevitz and Mansour, 1993)) *Consider any distribution D specified by a BN and its corresponding Fourier basis, and any Boolean function f. Algorithm KM$(D, f, \theta, \gamma, \delta)$ is given access to the BN representation of D and a MQ oracle for f and three accuracy parameters θ, γ, δ. The algorithm runs in time polynomial in n,$1/\theta$,$1/\gamma$,$1/\delta$ and with probability at least $1-\delta$ returns a list of sets $\mathcal{A} = \{\mathcal{S}\}$, estimates of the corresponding coefficients $\tilde{f}_\mathcal{S}$, and a hypothesis $h(x) = \sum_{\mathcal{S}\in\mathcal{A}} \tilde{f}_\mathcal{S}\phi_\mathcal{S}(x)$ such that (1) $\mathcal{A}$ includes all $\mathcal{S}$ such that $|\hat{f}_\mathcal{S}| \geq \theta$, (2) $|\mathcal{A}| \leq \frac{4}{\theta^2}$, (3) for all $\mathcal{S} \in \mathcal{A}$, $|\tilde{f}_\mathcal{S} - \hat{f}_\mathcal{S}| \leq \gamma$.*

As in prior works, the KM algorithm leads to learnability results for DNF. However, for this to hold, we need a bound on the spectral norm of conjunctions. The next few sections develop upper and lower bounds on the spectral norm.

5. Fourier Expansion of Conjunctions for k-Junta Distributions

In this section, we consider the class of k-junta distributions (Aliakbarpour et al., 2016) that model distributions where, intuitively, k of the n variables capture the complexity of the distribution. More formally, let generalized k-junta distributions be distributions such that conditioned on a set $J \subseteq [n]$ of size k, the remaining variables have a product distribution (uniform in the original definition). Similarly, a depth-d decision-tree (DT) distribution (Blanc et al., 2023) induces a uniform distribution in each leaf of the tree, hence it is a $k \leq 2^d$, k-junta distribution. Generalized k-junta distributions can be captured with a BN with a simple structure. Some arbitrary DAG represents the distribution over the set J, and all other variables depend on J and are conditionally independent given J. On the other hand, even chain BN provide flexibility that cannot be captured with shallow DT. For example, by the pigeonhole principle, a chain BN with n variables with significant correlation in conditional expectations cannot be captured with a DT distribution of depth $D < n/2 - 1$.

It is relatively easy to analyze the spectral norm of conjunctions for k-junta distributions. In particular, if f is a conjunction with d literals then its ancestor set includes at most $k + d$ variables. An argument as in Lemma 4 shows that the number of non-zero coefficients of f is bounded by 2^{k+d}. Then, due to the sparsity, we can use the argument for product distributions (Feldman, 2012) for the spectral norm. This yields the following lemma.

Lemma 8 *Let D be a k-junta distribution. Then the spectral norm of conjunctions f with d literals is bounded by $L_1(f) \leq 2^{(k+d)/2}$.*

6. Fourier Expansion of Conjunctions for Chain BN

In this section, we restrict our attention to linear chain BNs. These are later used as building blocks in the analysis of tree BNs. In a chain, each node has a single parent so that w.l.o.g. we can rename the variables so that the parent of any node i is node $i-1$.

Consider the conjunction $f(\mathbf{x}) := \bigwedge_{i\in\mathcal{T}_1} x_i \bigwedge_{j\in\mathcal{T}_0} \bar{x}_j$ where $\mathcal{T}_0$ and $\mathcal{T}_1$ are disjoint subsets of $[n]$. We introduce a series of notations using which we present a closed form expression for $\hat{f}_{\mathcal{S}}$. First, we use Φ_i to denote $\phi_i(\mathbf{X})$ which is a random variable depending on X_i and its parent X_{i-1}. Let $\mathcal{T} = \mathcal{T}_0 \cup \mathcal{T}_1$ and define the following random variable

$$\begin{aligned} Z_i =& X_i \mathbf{1}_{\{i\in\mathcal{T}_1\setminus\mathcal{S}\}} + (1-X_i)\mathbf{1}_{\{i\in\mathcal{T}_0\setminus\mathcal{S}\}} + \Phi_i \mathbf{1}_{\{i\in\mathcal{S}\setminus\mathcal{T}\}} \\ &+ \Phi_i X_i \mathbf{1}_{\{i\in\mathcal{S}\cap\mathcal{T}_1\}} + \Phi_i(1-X_i)\mathbf{1}_{\{i\in\mathcal{S}\cap\mathcal{T}_0\}} + \mathbf{1}_{\{i\notin\mathcal{T}\cup\mathcal{S}\}}. \end{aligned} \tag{3}$$

With this notation, it is not difficult to see that

$$\hat{f}_{\mathcal{S}} = \mathbb{E}_D[\phi_{\mathcal{S}}(\mathbf{X})f(\mathbf{X})] = \mathbb{E}_D\Big[\prod_{j\in[n]} Z_j\Big] = \mathbb{E}_{X_1}\Big[Z_1\mathbb{E}_{X_2}\Big[Z_2\cdots\mathbb{E}_{X_n}\Big[Z_n\Big|X_{n-1}\Big]\cdots\Big|X_1\Big]\Big], \tag{4}$$

where we used the fact that $X_1 \to \cdots \to X_n$ form a chain and that Z_i is a function of X_i, X_{i-1}. To analyze the expressions appearing in the iterative expectation note that $\mathbb{E}_{X_i}[X_i|X_{i-1}] = \mu_{i,X_{i-1}}$, and by Lemma 4, $\mathbb{E}_{X_i}[\Phi_i|X_{i-1}] = 0$. In addition, it is easy to verify that $\mathbb{E}_{X_i}[\Phi_i X_i|X_{i-1}] = \sigma_{i,X_{i-1}}$. This shows that in the iterative expectation, we may get a term with μ or with σ. The following definition provides the key to our analysis as it allows us to abstract the various cases in the same form, and by doing so to capture the combinatorial structure of parameters $\hat{f}_{\mathcal{S}}$:

Definition 9 (Recursive Form) *For any $i \in [n], x_{i-1} \in \{0,1\}$ and sets $\mathcal{T}_0, \mathcal{T}_1$ and $\mathcal{S}$ define*

$$\begin{aligned} A_i(x_{i-1}) =& \mu_{i,x_{i-1}}\mathbf{1}_{\{i\in\mathcal{T}_1\setminus\mathcal{S}\}} + (1-\mu_{i,x_{i-1}})\mathbf{1}_{\{i\in\mathcal{T}_0\setminus\mathcal{S}\}} + \sigma_{i,x_{i-1}}\mathbf{1}_{\{i\in\mathcal{S}\setminus\mathcal{T}\}} \\ &+ \sigma_{i,x_{i-1}}\mathbf{1}_{\{i\in\mathcal{S}\cap\mathcal{T}_1\}} - \sigma_{i,x_{i-1}}\mathbf{1}_{\{i\in\mathcal{S}\cap\mathcal{T}_0\}} + \mu_{i,x_{i-1}}\mathbf{1}_{\{i\notin\mathcal{T}\cup\mathcal{S}\}}. \end{aligned}$$

Moreover, with A_i we denote the random variable $A_i(X_{i-1})$ which is a function of X_{i-1}. Note that A_i satisfies the following identity

$$A_i = A_i(X_{i-1}) = A_i(0) + X_{i-1}(A_i(1) - A_i(0)). \tag{5}$$

It is not difficult to see that $\mathbb{E}_{X_i}[Z_i|X_{i-1}] = A_i$, when $i \in \mathcal{T}$. Hence we see that when evaluating (4) we may inherit a A_i term from the child and need to evaluate $\mathbb{E}[Z_{i-1}A_i]$. The following lemma develops a compact form for this expectation:

Lemma 10 *For any distribution for $X_{i-1}|X_{i-2}$, $\mathbb{E}_{X_{i-1}}[Z_{i-1}A_i|X_{i-2}] = b_i A_{i-1} + c_i$, where*

$$b_i = \begin{cases} A_i(1) & if\ i-1 \in \mathcal{T}_1 \\ A_i(0) & if\ i-1 \in \mathcal{T}_0 \\ A_i(1) - A_i(0) & if\ i-1 \notin \mathcal{T}_0 \cup \mathcal{T}_1 \end{cases} \quad \text{and} \quad c_i = \begin{cases} A_i(0) & if\ i-1 \notin \mathcal{S}\cup\mathcal{T} \\ 0 & otherwise \end{cases}$$

This allows us to recurse over all the iterative expectations to compute $\hat{f}_{\mathcal{S}}$. The additive term c_i yields hierarchically structured expressions. However, note that c_i will be canceled if the next expectation $j < i$ where $Z_j \neq 1$ is $j \in \mathcal{S}\setminus\mathcal{T}$. That is $\mathbb{E}[\Phi_j(c_i + b_i A_{i-1})] = \mathbb{E}[\Phi_j b_i A_{i-1}]$ because $\mathbb{E}[\Phi_j] = 0$. This is crucial in understanding the structure of $\hat{f}_{\mathcal{S}}$.

6.1. The Fourier Coefficients of Conjunctions

Definition 11 *For any $i \in [n]$ define $D_i := A_i(1) - A_i(0)$. Given any pair of subsets $\mathcal{S}$ and $\mathcal{T}$, let $a_1 < \cdots < a_m$ denote the the ordered elements of $\mathcal{S} \cup \mathcal{T}$. For $i \in [m]$ and any $r \in [a_{i-1}, a_i]$ define*

$$D'_{(r,a_i)} = \prod_{r \le \ell \le a_i} D_\ell, \qquad A'_{(r,a_i)}(0) = \sum_{r \le \ell \le a_i} D'_{(\ell+1,a_i)} A_\ell(0), \tag{6}$$

and by convention $D'_{(r,a_i)} = 1$ for any $r > a_i$. Let $A'_{(r,a_i)}(1) = A'_{(r,a_i)}(0) + D'_{(r,a_i)}$ and define the random variable

$$A'_{(r,a_i)} := A'_{(r,a_i)}(0) + X_{r-1}(A'_{(r,a_i)}(1) - A'_{(r,a_i)}(0)) = A'_{(r,a_i)}(0) + X_{r-1} D'_{(r,a_i)}.$$

Note that when $r > a_i$ then $A'_{(r,a_i)}(0) = 0$ and $A'_{(r,a_i)}(1) = D'_{(r,a_i)} = 1$.

These terms capture the hierarchical structure that arises from repeated application of Lemma 10. As a special case of the above definition $D'_{(a_i,a_i)} = D_{a_i}$ and $A'_{(a_i,a_i)} = A_{a_i}$. We highlight that $A', A'(1), A'(0)$ and D' satisfy the same relations as $A, A(1), A(0)$ and D. Moreover, we have:

Lemma 12 $A'_{(r,a_i)} = A'_{(r+1,a_i)}(0) + A_r D'_{(r+1,a_i)}$.

To develop the expression for $\hat{f}_{\mathcal{S}}$ we split the chain into segments and analyze the expectation over segments. The next two lemmas prepare the ground by analyzing individual segments with boundaries at T nodes.

Lemma 13 (Single $\mathcal{S} \cup \mathcal{T}$ segment) *Consider a chain of nodes $X_0 \to \cdots \to X_{n+1}$ that form a segment in the sense that $\mathcal{S}, \mathcal{T} \subseteq \{1\}$. Then,*

$$\begin{aligned}\mathbb{E}_{X_1,\ldots,X_n}\left[Z_1 A'_{(n+1,n+1)}(X_n)|X_0\right] =& A'_{(1,n+1)}(X_0)\mathbf{1}_{\{1 \notin \mathcal{T} \cup \mathcal{S}\}} + D'_{(2,n+1)} A_1(X_0)\mathbf{1}_{\{1 \in \mathcal{S} \setminus \mathcal{T}\}} \\ &+ A'_{(2,n+1)}(1) A_1(X_0)\mathbf{1}_{\{1 \in \mathcal{T}_1\}} + A'_{(2,n+1)}(0) A_1(X_0)\mathbf{1}_{\{1 \in \mathcal{T}_0\}}.\end{aligned}$$

Proof Assume $n > 1$ as for $n = 1$ the lemma follows from Lemma 10. By the law of total expectation, the expectation in the lemma iterates as $\mathbb{E}_{X_1}\left[Z_1 \mathbb{E}_{X_2}\left[\cdots \mathbb{E}_{X_n}\left[A'_{(n+1,n+1)}\middle|X_{n-1}\right]. \cdots \middle|X_1\right]\middle|X_0\right]$. By an inductive argument we prove that

$$\mathbb{E}_{X_2}\left[\cdots \mathbb{E}_{X_n}\left[A'_{(n+1,n+1)}\middle|X_{n-1}\right]. \cdots \middle|X_1\right] = A'_{(2,n+1)}.$$

First, observe that $A'_{(n+1,n+1)}(X_n) = A_{n+1}(X_n)$. From Lemma 10, as $n \notin \mathcal{S} \cup \mathcal{T}$, the innermost expectation equals $\mathbb{E}_{X_n}\left[A'_{(n+1,n+1)}|X_{n-1}\right] = D_{n+1}A_n + A_{n+1}(0) = A'_{(n,n+1)}$, where the last equality follows from Lemma 12. Assuming that $\mathbb{E}_{X_m,\cdots,X_n}\left[A'_{(n+1,n+1)}\middle|X_{m-1}\right] = A'_{(m,n+1)}$ holds for some $2 < m \le n$, then for $(m-1)$ we have

$$\begin{aligned}\mathbb{E}_{X_{m-1},\cdots,X_n}\left[A'_{(n+1,n+1)}\middle|X_{m-2}\right] &= \mathbb{E}_{X_{m-1}}\left[A'_{(m,n+1)}|X_{m-2}\right] \\ &= A'_{(m,n+1)}(0) + D'_{(m,n+1)}\mathbb{E}_{X_{m-1}}[X_{m-1}|X_{m-2}] \\ &= A'_{(m,n+1)}(0) + D'_{(m,n+1)}A_{m-1} = A'_{(m-1,n+1)},\end{aligned}$$

where the last line holds because $m-1 \notin \mathcal{S} \cup \mathcal{T}$ and $\mu_{m-1,X_{m-2}} = A_{m-1}$ and due to Lemma 12. This established the induction. Lastly, the outermost expectation equals

$$\mathbb{E}_{X_1}\left[Z_1 A'_{(2,n+1)} \middle| X_0\right] = A'_{(2,n+1)}(0)\mathbb{E}[Z_1|X_0] + D'_{(2,n+1)}\mathbb{E}_{X_1}[Z_1 A_2|X_0].$$

The rest of the proof follows by evaluating $\mathbb{E}[Z_1|X_0]$ and $\mathbb{E}[Z_1 A_2|X_0]$ and simplifying using Lemma 10 and 12. ■

The lemma implies an alternative definition of A' that is insightful:

$$A'_{(r,a_i)}(X_{r-1}) = \mathbb{E}_{X_r,\ldots,X_{a_i-1}}[A_{a_i}|X_{r-1}]. \tag{7}$$

Lemma 14 (Single $\mathcal{T}$-segment, multiple $\mathcal{S}$-segments) *Consider a chain $X_0 \to \cdots \to X_{n+1}$ that potentially has multiple $\mathcal{S}$ nodes inside a $\mathcal{T}$ segment, where $\mathcal{T} \subseteq \{1\}$ and $\mathcal{S} \subseteq [n]$. If $\mathcal{S} \setminus \mathcal{T}$ is not empty, define $s_\circ := \min \mathcal{S} \setminus \mathcal{T}$; otherwise $s_\circ = n+1$. Let $\Gamma = \mathbb{E}\left[\prod_{1\le j\le n} Z_j A'_{(n+1,n+1)} \middle| X_0\right]$. When $\mathcal{T} = \emptyset$, $\Gamma = D'_{(s_\circ+1,n+1)} A'_{(1,s_\circ)}(X_0)$; when $\mathcal{T} = \{1\}$, $\Gamma = D'_{(s_\circ+1,n+1)} A'_{(2,s_\circ)}(y_1) A_1(X_0)$, where $y_n = \mathbf{1}_{(n\in\mathcal{T}_1)}$, and $y_1 = \mathbf{1}_{(1\in\mathcal{T}_1)}$.*

Proof Sketch: The proof follows by law of iterative expectations and by breaking the chain into segments and applying Lemma 13 on each segment. Let Γ be the expectation of interest as in the lemma's statement. If $\mathcal{S} \setminus \mathcal{T}$ is not empty, let $s_\circ = a_1 < \cdots < a_k$ be the ordered elements of $\mathcal{S} \setminus \mathcal{T}$ with $k = |\mathcal{S} \setminus \mathcal{T}|$. We consider two cases depending on $\mathcal{T}$ elements. For compactness, we provide the proof only for the first case as the proof steps are similar.

Case (a) ($\mathcal{T} = \emptyset$): In this case, the segments are $[1, a_1)$, $[a_i, a_{i+1})$ for $i = 1, \ldots, k-1$, and $[a_k, n+1)$. Starting from the tail segment $[a_k, n+1)$, from Lemma 13, the contribution is

$$\Gamma_{[a_k,n+1)} = \mathbb{E}_{X_{a_k},\ldots,X_n}\left[Z_{a_k} A'_{(n+1,n+1)}(X_n)|X_{a_k-1}\right] = D'_{(a_k+1,n+1)} A_{a_k}(X_{a_k-1}).$$

Note that $A_{a_k} = A'_{(a_k,a_k)}$, implying that the input to the next segment is an A' term. From this point for any $\mathcal{S}$-segment $[a_i, a_{i+1})$, the contribution is $\Gamma_{[a_i,a_{i+1})} = D'_{(a_i+1,a_{i+1})} A'_{(a_i,a_i)}$. Continue this argument until the head segment, if $s_\circ = 1$, the head segment is $[a_1, a_2)$ that has been covered and contributing $A'_{(1,1)}$. If $s_\circ > 1$, then the head segment is $[1, s_\circ)$. The input to this segment is $A'_{(a_1,a_1)}$. Again from Lemma 13, as $1 \notin \mathcal{S} \cup \mathcal{T}$, the contribution is $\Gamma_{[1,s_\circ)} = \mathbb{E}_{X_1,\ldots,X_{s_\circ-1}}\left[Z_1 A'_{(s_\circ,s_\circ)}|X_0\right] = A'_{(1,s_\circ)}(X_0)$. Lastly, multiplying all the contributions gives the desired expression, by noting that products of consecutive D' terms gives another D' with the total interval. Now consider $\mathcal{S} \setminus \mathcal{T} = \emptyset$ which means $s_\circ = n+1$. Since $\mathcal{T}$ is empty then so is $\mathcal{S}$. Hence, we have an empty chain and from Lemma 13, the contribution is $A'_{(1,n+1)}(X_0)$ which is consistent with the definition of Γ. ■

The lemma implies that $\mathcal{S}$ segments inside a $\mathcal{T}$ segment contribute a D' term which is a product of D's from the the minimum $\mathcal{S}$-only node to the end $\mathcal{T}$ node. We are finally ready to express $\hat{f}_\mathcal{S}$.

Lemma 15 (Fourier coefficients of conjunctions) *Consider a conjunction f with $\mathcal{T}$ being the set of literals. Let $\mathcal{T}_0 \subset \mathcal{T}$ be the negated literals and $\mathcal{T}_1 = \mathcal{T} \setminus \mathcal{T}_0$ be the set of literals without negation. For any $\mathcal{S} \subseteq [n]$, if $\max \mathcal{S} > \max \mathcal{T}$, then $\hat{f}_\mathcal{S} = 0$; otherwise*

$$\hat{f}_\mathcal{S} = \prod_{a_j \in \mathcal{T}_0 \cup \{0\}} A'_{(a_j+1,a_{j+1})}(0) \prod_{a_k \in \mathcal{T}_1} A'_{(a_k+1,a_{k+1})}(1) \prod_{a_l \in \mathcal{S}\setminus\mathcal{T}} D'_{(a_l+1,a_{l+1})},$$

where $0 < a_1 < a_2 < \cdots$ are the ordered elements of $\mathcal{S} \cup \mathcal{T} \cup \{0\}$.

Proof Sketch: The main idea is to break the chain into the $\mathcal{T}$ segments and use Lemma 14 repeatedly. Let $t^* = \max(\mathcal{T})$ and $s^* = \max(\mathcal{S})$. When $s^* > t^*$, $\hat{f}_{\mathcal{S}} = 0$, because the only term dependent on X_{s^*} is Φ_{s^*}, and the inner most expectation equals $\mathbb{E}_{X_{s^*}}\left[Z_{s^*}\middle|X_{s^*-1}\right] = \mathbb{E}_{X_{s^*}}\left[\Phi_{s^*}\middle|X_{s^*-1}\right] = 0$, where we used Lemma 4 for the last equality. Suppose $s^* \leq t^*$, and let $t_1 < \cdots, < t_d = t^*$ be the ordered elements of $\mathcal{T}$ with $d = |\mathcal{T}|$. These nodes create $d-1$, $\mathcal{T}$-segments $[t_{i-1}, t_i)$, a head segment $[1, t_1)$ and a tail segment $[t^*, n]$. Let $t_0 := 0$, $\mathcal{S}_i := \mathcal{S} \cap [t_{i-1}, t_i)$ for any $i \in [d]$ and note that since $\mathcal{S} \subseteq [n]$, we have $\mathcal{S}_1 \subseteq [1, t_1)$. If $\mathcal{S}_i \setminus \mathcal{T}$ is not empty, define $h_i := \min \mathcal{S}_i \setminus \mathcal{T}$; otherwise set $h_i = t_i$. For any $i \leq j$, let $\mathbf{Z}_i^j = \prod_{i \leq l \leq j} Z_l$. Then, $\hat{f}_{\mathcal{S}}$ breaks into expectations over each segment:

$$\hat{f}_{\mathcal{S}} = \mathbb{E}_{[X_1, X_{t_1})}\left[\mathbf{Z}_1^{t_1-1} \cdots \mathbb{E}_{[X_{t_{d-1}}, X_{t^*})}\left[\mathbf{Z}_{t_{d-1}}^{t^*-1}\mathbb{E}_{[X_{t^*}, n]}\left[Z_{t^*}\middle|X_{t^*-1}\right]\middle|X_{t_{d-1}-1}\right] \cdots \middle|X_0\right].$$

Starting from the tail, we use an inductive argument to calculate $\hat{f}_{\mathcal{S}}$. By repeated use of Lemma 14, we show that each segment $[t_{i-1}, t_i)$ inherits $A'_{(t_i,t_i)}(X_{t_i-1})$ as the input from the previous inner segment and contributes a variable $A'_{(t_{i-1},t_{i-1})}(X_{t_{i-1}-1})$ multiplied by a constant which is comprised of a product of D' and A' terms as in Lemma 14. As a result, $\hat{f}_{\mathcal{S}}$ is the product of all the constants. Lastly, multiplying the constants produced by the segments and letting $y_0 = 0$ gives:

$$\hat{f}_{\mathcal{S}} = \prod_{i=1}^{d} D'_{(h_i+1,t_i)} A'_{(t_{i-1}+1,h_i)}(y_{i-1}). \tag{8}$$

■

6.2. Upper Bound on the Spectral Norm of Conjunctions Under Chain BNs

Lemma 16 (Bounding A' terms) *If D is a c-bounded chain, with $c \in (0, \frac{1}{2})$, then, $\left|A'_{(r,a_i)}(0)\right| \leq 1 - c$ and $\left|A'_{(r,a_i)}(1)\right| \leq 1 - c$, for any $a_i \in \mathcal{S} \cup \mathcal{T}$ and $r \in [a_{i-1}, a_i]$.*

Theorem 17 *Suppose D is a c-bounded chain with $c \in (0, \frac{1}{2})$, and $|D_{\sigma,i}| = |\sigma_{i,1} - \sigma_{i,0}| \leq D_\sigma$, and $|D_{\mu,i}| = |\mu_{i,1} - \mu_{i,0}| \leq D_\mu$ with $D_\sigma + D_\mu = 2\alpha < 1$. Then, for any conjunction f with d literals $L_1(f) \leq \left(\frac{(2-2\alpha)(1-c)}{1-2\alpha}\right)^d$.*

Proof Sketch: Recall the notations $A'_{(r,a_i)}, D'_{(r,a_i)}, a_i, t_i, \mathcal{S}_i, h_i, t^*, s^*$ as in Lemma 15. From (8) in Lemma 15, we have the expression for $\hat{f}_{\mathcal{S}}$. By Lemma 16 and the theorem's assumption,

$$|\hat{f}_{\mathcal{S}}| \leq (1-c)^d \prod_{i \in [d]} \left|D'_{(h_i+1,t_i)}\right|.$$

If $s^* > t^*$, then $\hat{f}_{\mathcal{S}} = 0$. Therefore, summing over $|\hat{f}_{\mathcal{S}}|$ for all $\mathcal{S} \subseteq [t^*]$ gives the L_1 bound

$$L_1(f) = \sum_{\mathcal{S} \subseteq [t^*]} \left|\hat{f}_{\mathcal{S}}\right| \leq (1-c)^d \sum_{\mathcal{S} \subseteq [t^*]} \prod_i \left|D'_{(h_i+1,t_i)}\right|.$$

By definition, $\cup_i \mathcal{S}_i$ covers the set $\{1, \cdots, t^* - 1\}$. Therefore, we can break the summation into summations over each $\mathcal{S}_i$. Then, we show (detailed skipped) that the summations can be interchanged with the product. Hence, L1 bound equals $(1-c)^d \prod_{i \in [d]} \left(\sum_{\mathcal{S}'_i \subseteq (t_{i-1}, t_i]} \left|D'_{(h_i+1,t_i)}\right|\right)$. This shows a decoupling phenomenon, where $L_1(f)$ is related to the product of L_1 norms on each segment.

When $h_i = k < t_i$, and $|\mathcal{S}_i| = \ell$, then $\left|D'_{(h_i+1,t_i)}\right| \leq D_\sigma^{\ell-1} D_\mu^{t_i-k-\ell+1}$, where we used Definition 9 implying that we get $D_{\sigma,j}$ when $j \in \mathcal{S}$, and $D_{\mu,j}$ otherwise. Moreover, by a combinatorial argument, and using the binomial theorem, we upper bound each L_1 norm by

$$1 + \sum_{k=t_{i-1}+1}^{t_i} \sum_{r=0}^{t_i-k} \binom{t_i-k}{r} D_\sigma^r D_\mu^{t_i-k-r} = 1 + \sum_{k'=0}^{t_i-t_{i-1}-1} (D_\sigma + D_\mu)^{k'} \leq \frac{2 - D_\sigma - D_\mu}{1 - D_\sigma - D_\mu},$$

where the last inequality holds by increasing the range of k' to ∞. Now combining the bounds gives the desired result. ■

The bound in Theorem 17 can be made tighter by a slightly more refined analysis that accounts for A' terms explicitly instead of bounding them by $1 - c$. This is especially useful for product distributions that can be viewed as a chain (with an arbitrary ordering) where for all i we have $\mu_{i,0} = \mu_{i,1} = \mu_i$. Here we obtain an exact characterization and improve over the $(\sqrt{2})^d$ bound:

Proposition 18 *The spectral norm of a conjunction f of d literals in product distributions equals $L_1(f) = (\prod_{i\in\mathcal{T}_1}(\mu_i + \sigma_i))(\prod_{j\in\mathcal{T}_0}((1-\mu_j) + \sigma_j)) \leq 1.21^d$.*

7. Fourier Expansion of Conjunctions for Tree BNs

Theorem 19 *For any α-difference bounded tree BN, the spectral norm of any conjunction f with d literals is bounded by $L_1(f) \leq \left(\frac{2-2\alpha}{1-2\alpha}\right)^{2d}$.*

Proof Sketch: The proof builds upon a decoupling argument by viewing the BN tree as a collection of connected branches (chains). More precisely, a branch is a set of nodes that form a chain starting from an expansion node and ending with a leaf or another expansion node. Our proof shows that $L_1(f)$ decomposes into the product of L_1 norms related to each branch. However, unlike chains, the major difficulty for trees is that a branch can have multiple children that produce related A' terms. Hence, the expectation in a generic branch takes the product of multiple A' terms as input. To address this we introduce the concept of single-sided bounded functions. A function $f : \{0,1\} \to \mathbb{R}$ is said to be *bounded single-sided* if $\max_x |f(x)| \leq 1$ and $f(0)$ and $f(1)$ have the same sign. We need the following generalized analysis for chains, which is proved similarly to Lemma 15 and 17.

Lemma 20 (Generic branch) *For any α-difference bounded chain, a bounded single-sided f, and for any $\mathcal{S}, \mathcal{T} \subseteq [n]$ and the corresponding $Z_i, i \in [n]$, $\mathbb{E}\left[\prod_{i\in[n]} Z_i \;\; f(X_n)\middle| X_0\right] = b'_{\mathcal{S},\mathcal{T}}\, g(X_0)$, for some bounded single-sided g and $b'_{\mathcal{S},\mathcal{T}}$ with L_1 bound $\sum_{\mathcal{S}} \left|b'_{\mathcal{S},\mathcal{T}}\right| \leq \left(1 + \frac{|f(1)-f(0)|}{1-2\alpha}\right)\left(\frac{2-2\alpha}{1-2\alpha}\right)^{|\mathcal{T}|}$.*

We proceed by induction in a reverse topological order starting from leaf branches to the root branch, showing that each branch contributes a constant $b'_{\mathcal{C}_i}$ and outputs an A'-type term to the parent branch. From Lemma 20, the contribution of the leaf branch $\mathcal{C}_i$ is $\mathbb{E}_{\sim\mathcal{C}_i}\left[Z_{\mathcal{C}_i}\middle| X_{\mathrm{pa}(\mathcal{C}_i)}\right] = b'_{\mathcal{C}_i} g_k(X_{\mathrm{pa}(\mathcal{C}_i)})$, For the inductive step, assume the claim holds for all descendants of branch $\mathcal{C}_i$. Because the child branches are independent of each other conditioned on the parent, their A' terms will be multiplied to create the input to the parent branch. Let $\mathcal{C}_{i_1}, \cdots, \mathcal{C}_{i_{m_i}}$ be the child branches of $\mathcal{C}_i$, where m_i is the number of child branches. The contribution of each child-branch $\mathcal{C}_{i_j}$ is $b'_{\mathcal{C}_{i_j}} g_{i_j}(X_{\mathrm{pa}(\mathcal{C}_{i_j})})$. The constants b' can be taken out of the expectation. Each g_{i_j} is a function of the last element in $\mathcal{C}_i$, which is an expansion node denoted by e_i. Hence, the contribution of $\mathcal{C}_i$ is calculated as $\mathbb{E}_{\sim\mathcal{C}_i}\left[Z_{\mathcal{C}_i} \prod_j g_{i_j}(X_{e_i})\middle| X_{\mathrm{pa}(\mathcal{C}_i)}\right]$. To handle the product we have:

Lemma 21 *Finite product $\prod_i f_i$ of bounded single-sided functions is bounded single-sided.*

Therefore, the branch $\mathcal{C}_i$ takes as input a single-sided function $f = \prod_j g_{i_j}$ which is an A'-type term. Then, using Lemma 20, this branch produces a bounded single-sided function $g_i(X_{\mathrm{pa}(\mathcal{C}_i)})$ multiplied by a constant $b'_{\mathcal{C}_i}$. With that the induction is established meaning that any branch $\mathcal{C}_i$ contributes the constant $b'_{\mathcal{C}_i}$ given as in Lemma 20. As a result, $\hat{f}_{\mathcal{S}} = \prod_{i=0}^{k} b'_{\mathcal{C}_i}\, g_0(0)$. Then using the decoupling argument, we show that $L_1(f) \leq \prod_{i=0}^{k}(\sum_{\mathcal{S}\cap\mathcal{C}_i} |b'_{\mathcal{C}_i}|)$, The L1 bound for any leaf branch is $\left(\frac{2-2\alpha}{1-2\alpha}\right)^{|\mathcal{T}\cap\mathcal{C}_i|}$, while it is $\left(\frac{2-2\alpha}{1-2\alpha}\right)^{|\mathcal{T}\cap\mathcal{C}_i|+1}$ for non-leaf branches. The $+1$ in the exponent is from Lemma 20, and the fact that $1 + \frac{|f(1)-f(0)|}{1-2\alpha} \leq \frac{2-2\alpha}{1-2\alpha}$. Multiplying these bounds, the L1 bound for the tree is $\left(\frac{2-2\alpha}{1-2\alpha}\right)^{|\mathcal{T}|+E}$ where E is the number of non-leaf branches. We can ignore leaf branches without any $\mathcal{T}$ nodes. Hence the number of leaf branches is $\leq d$ and since we have a tree the number of non-leaf branches is at most $d-1$ and $E \leq d$ which gives the desired L1 bound. ■

8. Lower Bounds on the Spectral Norm

Our upper bounds for chains and trees require $D_\mu + D_\sigma < 1$. In this section, we show that this is necessary. Without this condition, even for chains the spectral norm of a single literal can be exponentially large. For general graphs, the L1 can be exponentially large even when $D_\mu + D_\sigma < 1$.

Lemma 22 (Lower bound for chains) *Consider a chain BN with $n+2$ variables $X_0, \ldots, X_{n+1}$ and the function $f = X_{n+1}$. Then (1) There exist distributions instantiating this structure where all nodes share the same conditional distribution table, with $D_\mu = |\mu_{i,1} - \mu_{i,0}| > 0$, $D_\sigma = |\sigma_{i,1} - \sigma_{i,0}| > 0$ and $L_1(f) = \Omega((D_\mu + D_\sigma)^n)$. (2) For any $c \leq 0.0246$, choosing $\mu_1 = c$, $\mu_2 = 0.5 + c + 2\sqrt{c}$ (with $D_\mu = 0.5 + 2\sqrt{c}$) yields $D_\mu + D_\sigma > 1 + c$ and $L_1(f) = \Omega((1+c)^n)$.*

The construction for general graphs is significantly more complex and due to space constraints we only sketch some of the ideas. The main idea is that if nodes can have multiple parents then the computation of their coefficients can generate a structure similar to conjunctions of large d and hence have an exponentially large spectral norm. We construct a graph G^* as follows: G^* includes n independent chains each with 23 nodes. In addition, the leaf of each chain is one of the roots of a complete anti-binary tree (where each node has two parents and one child) with n roots and one leaf. Denote the root nodes of the anti-tree by $X_1, \ldots, X_n$. For node i with parents j, k in the anti-tree we use a conditional probability of the form $p(X_i = 1 | X_j X_k = (00, 01, 10, 11)) = (\alpha + D, \alpha, \alpha, \alpha + D)$. Analyzing the coefficients we show that the anti-tree portion yields $(2D)^{n-1} X_1 \ldots X_n$, i.e., an exponentially small constant multiplied by a conjunction with n variables. We then show that the contribution from the conjunction dominates $(2D)^{n-1}$. This yields:

Theorem 23 (Exponential Lower Bound for General Graphs) *Consider the function $f = V_1$, where V_1 is the unique leaf of the tree in G^*, i.e., a conjunction of size 1. There exist bounded BN distributions defined by G^* such that for all nodes i, $\mu_{i,pa(i)} \geq 0.01$ and for all assignments γ_1, γ_2 to the parents of node i, $|\mu_{i,pa(i)=\gamma_1} - \mu_{i,pa(i)=\gamma_2}| < 0.49$ and $L_1(f) = \Omega(1.05^n)$.*

9. Implications for DNF Learnability

Given the analysis of the KM algorithm and bounds on the spectral norm, results previously proved for the uniform or product distributions, hold more or less directly in the new setting. A key re-

quirement in this analysis is that the spectral norm of conjunctions is bounded. In this section we assume that such a bound exists. In particular for a distribution D and conjunction g with d literals, we assume that $L_1(g) \leq L_1(d)$ for some function $L_1(d)$ and present the results in this general form. Hence learnability holds whenever $L_1(d)$ is available, including for difference bounded tree distributions where $L_1(d) = O((\frac{2}{1-2\alpha})^{2d})$, and k-junta distributions where $L_1(d) = O(2^{(k+d)/2})$.

Recall that disjoint DNFs are disjunctions of conjunctions where the conjunctions are mutually exclusive and decision trees whose node tests are individual variables are a subset of disjoint DNF. With the bound for conjunctions and the KM algorithm, we can follow previous work to show:

Corollary 24 (cf. Theorem 3.10 of (Kushilevitz and Mansour, 1993)) *Consider any distribution D with its corresponding Fourier basis where $L_1(d)$ is a bound on the spectral norm of conjunctions of size d. Let f be any s-term disjoint DNF and let $h(x)$ be the output of KM$(D, f, \theta, \gamma, \delta)$ with $\theta = \epsilon/2L_1$ and $\gamma = \epsilon^3/16L_1^2$ where $L_1 = sL_1(d)$ and $d = \log_{1-c}(\epsilon/4s)$. Then with probability at least $1-\delta$, $P_D(f(X) \neq \text{sign}(h(X))) \leq \epsilon$.*

Mansour (1995) has shown that the KM algorithm cannot be used directly to learn (non-disjoint) DNF. Despite this, two approaches exist to learn DNF through the Fourier basis. The first by Jackson (1997) uses boosting to learn a Fourier based representation (but where the coefficients are different from the coefficients of f). The second, by Feldman (2012), gives a more direct algorithm with the same effect. We observe that with small modifications Feldman's algorithm can be used directly with our basis. In particular, Feldman (2012)'s analysis and algorithm exclude high-degree coefficients due to sparsity. However, this is not required for the correctness or efficiency of the algorithm. Following these observations we can show:

Corollary 25 (cf. Corollary 15 in (Feldman, 2012)) *Consider any distribution D with its corresponding Fourier basis where $L_1(d)$ is a bound on the spectral norm of conjunctions of size d. Let f be any s-term DNF and let $g(x)$ be the output of PTFconstruct(D, f, γ, δ) where $\epsilon' = \epsilon/6$, $d = \log_{1-c}(\epsilon'/4s)$, $L_1 = 2sL_1(d) + 1$, and $\gamma = \frac{\epsilon'}{L_1}$. Then with probability at least $1-\delta$, $P_D(f(X) \neq g(X)) \leq \epsilon$.*

10. Conclusion

The paper develops a generalized Fourier basis and shows that major algorithmic tools from learning theory can be used with this basis. We emphasize that the basis and the extended KM algorithm are valid for any distribution. Learnability of DNF was established for generalized k-junta distributions and for difference bounded tree distributions by bounding the spectral norm of conjunctions in these cases. The introduction of the generalized Fourier basis suggests many questions for future work. These include understanding what classes of graphs and constraints on parameters have bounded spectral norm, the potential of the low degree algorithm for constant depth circuits (Linial et al., 1993) with the new basis, and whether some of the applications of the Fourier representation for the uniform distribution (see (O'Donnell, 2014; Wolf, 2008)) can be generalized to the new basis.

Acknowledgments

This work was partially supported by the NSF Grant CCF-2211423.

References

Howard Aizenstein, Avrim Blum, Roni Khardon, Eyal Kushilevitz, Leonard Pitt, and Dan Roth. On learning read-k-satisfy-j DNF. *SIAM Journal on Computing*, 27(6):1515–1530, 1998.

Maryam Aliakbarpour, Eric Blais, and Ronitt Rubinfeld. Learning and testing junta distributions. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *Proceedings of the 29th Conference on Learning Theory*, volume 49, pages 19–46, 2016.

Dana Angluin and Michael Kharitonov. When won't membership queries help? *Journal of Computer and System Sciences*, 50(2):336–355, 1995.

M. Bellare. The spectral norm of finite functions. Technical report, MIT, USA, 1991.

Arnab Bhattacharyya, Sutanu Gayen, Eric Price, Vincent Y. F. Tan, and N. V. Vinodchandran. Near-optimal learning of tree-structured distributions by Chow and Liu. *SIAM Journal of Comput.*, 52 (3):761–793, 2023.

Guy Blanc, Jane Lange, Ali Malik, and Li-Yang Tan. Lifting uniform learners via distributional decomposition. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1755–1767. ACM, 2023.

Avrim Blum, Merrick L. Furst, Jeffrey C. Jackson, Michael J. 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*, pages 253–262. ACM, 1994.

Nader H. Bshouty. Exact learning boolean function via the monotone theory. *Information and Computation*, 123(1):146–153, 1995.

Nader H. Bshouty and Christino Tamon. On the Fourier spectrum of monotone functions. *Journal of the ACM*, 43(4):747–770, 1996.

Nader H. Bshouty, Jeffrey C. Jackson, and Christino Tamon. More efficient PAC-learning of DNF with membership queries under the uniform distribution. *Journal of Computer and System Sciences*, 68(1):205–234, 2004.

C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, May 1968.

Vitaly Feldman. Attribute-efficient and non-adaptive learning of parities and DNF expressions. *Journal of Machine Learning Research*, 8:1431–1460, 2007.

Vitaly Feldman. Learning DNF expressions from Fourier spectrum. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 17.1–17.19, 2012.

Merrick L Furst, Jeffrey C Jackson, and Sean W Smith. Improved learning of AC^0 functions. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, pages 317–325, 1991.

O. Goldreich and L. A. Levin. A hard-core predicate for all one-way functions. In *Proceedings of the twenty-first annual ACM symposium on Theory of computing*, pages 25–32. ACM Press, 1989.

M. Heidari and R. Khardon. Learning DNF through generalized Fourier representations. *arXiv preprint 2506.01075*, 2025.

Mohsen Heidari, Jithin Sreedharan, Gil I Shamir, and Wojciech Szpankowski. Finding relevant information via a discrete Fourier expansion. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4181–4191, 2021.

Mohsen Heidari, Jithin K. Sreedharan, Gil Shamir, and Wojciech Szpankowski. Sufficiently informative and relevant features: An information-theoretic and Fourier-based characterization. *IEEE Transactions on Information Theory*, 68(9):6063–6077, September 2022.

Lisa Hellerstein, Devorah Kletenik, Linda Sellie, and Rocco A. Servedio. Tight bounds on proper equivalence query learning of DNF. In *The 25th Annual Conference on Learning Theory*, pages 31.1–31.18, 2012.

Klaus-Uwe Höffgen. Learning and robust learning of product distributions. In Lenny Pitt, editor, *Proceedings of the Sixth Annual ACM Conference on Computational Learning Theory*, pages 77–83. ACM, 1993.

Jeffrey C Jackson. An efficient membership-query algorithm for learning DNF with respect to the uniform distribution. *Journal of Computer and System Sciences*, 55(3):414–440, December 1997.

Adam Tauman Kalai, Alex Samorodnitsky, and Shang-Hua Teng. Learning and smoothed analysis. In *2009 50th Annual IEEE Symposium on Foundations of Computer Science*, pages 395–404. IEEE, October 2009.

Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

Roni Khardon. On using the Fourier transform to learn disjoint DNF. *Information Processing Letters*, 49(5):219–222, 1994.

Adam R. Klivans and Rocco A. Servedio. Learning DNF in time $2^{\tilde{O}(n^{1/3})}$. *Journal of Computer and System Sciences*, 68(2):303–318, 2004.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models - Principles and Techniques*. MIT Press, 2009.

Eyal Kushilevitz. A simple algorithm for learning O(log n)-term DNF. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, pages 266–269, 1996.

Eyal Kushilevitz and Yishay Mansour. Learning decision trees using the Fourier spectrum. *SIAM Journal on Computing*, 22(6):1331–1348, December 1993.

Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, Fourier transform, and learnability. *J. ACM*, 40(3):607–620, 1993.

Y. Mansour. An O(n log log n) learning algorithm for DNF under the uniform distribution. *Journal of Computer and System Sciences*, 50(3):543–550, June 1995.

Ryan O'Donnell. *Analysis of Boolean functions*. Cambridge University Press, 2014.

Judea Pearl. *Probabilistic reasoning in intelligent systems - networks of plausible inference*. Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann, 1989.

Yoshifumi Sakai and Akira Maruoka. Learning monotone log-term DNF formulas under the uniform distribution. *Theory of Computing Systems*, 33(1):17–33, 2000.

Rocco A. Servedio. On learning monotone DNF under product distributions. *Information and Computation*, 193(1):57–74, 2004.

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, November 1984.

Leslie G. Valiant. Learning disjunction of conjunctions. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, pages 560–566. Morgan Kaufmann, 1985.

Karsten A. Verbeurgt. Learning DNF under the uniform distribution in quasi-polynomial time. In *Proceedings of the Third Annual Workshop on Computational Learning Theory, COLT*, pages 314–326, 1990.

Ronald de Wolf. *A Brief Introduction to Fourier Analysis on the Boolean Cube*. Number 1 in Graduate Surveys. Theory of Computing Library, 2008.