Mansour's Conjecture is True for Random DNF Formulas

Adam Klivans
University of Texas at Austin
klivans@cs.utexas.edu

Homin K. Lee
University of Texas at Austin
homin@cs.utexas.edu

Andrew Wan
Columbia University
atw12@cs.columbia.edu

Abstract

In 1994, Y. Mansour conjectured that for every DNF formula on n variables with t terms there exists a polynomial p with $t^{O(\log(1/\epsilon))}$ non-zero coefficients such that $\mathbf{E}_{x\in\{0,1\}^n}[(p(x) - f(x))^2] \leq \epsilon$. We make the first progress on this conjecture and show that it is true for several natural subclasses of DNF formulas including randomly chosen DNF formulas and read-k DNF formulas for constant k.

Our result yields the first polynomial-time query algorithm for agnostically learning these subclasses of DNF formulas with respect to the uniform distribution on $\{0,1\}^n$ (for any constant error parameter).

Applying recent work on sandwiching polynomials, our results imply that a $t^{-O(\log 1/\epsilon)}$-biased distribution fools the above subclasses of DNF formulas. This gives pseudorandom generators for these subclasses with shorter seed length than all previous work.

1 Introduction

Let $f : \{0,1\}^n \to \{0,1\}$ be a DNF formula, *i.e.,* a function of the form $T_1 \vee \cdots \vee T_t$ where each T_i is a conjunction of at most n literals. In this paper we are concerned with the following question: How well can a real-valued polynomial p approximate the Boolean function f? This is an important problem in computational learning theory, as real-valued polynomials play a critical role in developing learning algorithms for DNF formulas.

Over the last twenty years, considerable work has gone into finding polynomials p with certain properties (*e.g.,* low-degree, sparse) such that

$$\mathop{\mathbf{E}}_{x\in\{0,1\}^n}[(p(x) - f(x))^2] \leq \epsilon.$$

In 1989, Linial et al. (1993) were the first to prove that for any t-term DNF formula f, there exists a polynomial $p : \{0,1\}^n \to \mathbb{R}$ of degree $O(\log(t/\epsilon)^2)$ such that $\mathbf{E}_{x\in\{0,1\}^n}[(p(x) - f(x))^2] \leq \epsilon$. They showed that this type of approximation implies a quasipolynomial-time algorithm for PAC learning DNF formulas with respect to the uniform distribution. Kalai et al. (2008) observed that this fact actually implies something stronger, namely a quasipolynomial-time agnostic learning algorithm for learning DNF formulas (with respect to the uniform distribution). Additionally, the above approximation was used in recent work due to Bazzi (2007) and Razborov (2008) to show that bounded independence fools DNF formulas.

Three years later, building on the work of Linial et al. (1993) Mansour (1995) proved that for any t-term DNF formula, there exists a polynomial p defined over $\{0,1\}^n$ with *sparsity* $t^{O(\log\log t \log(1/\epsilon))}$ such that $\mathbf{E}_{x\in\{0,1\}^n}[(p(x) - f(x))^2] \leq \epsilon$ (for $1/\epsilon = \mathrm{poly}(n)$). By sparsity we mean the number of non-zero Fourier coefficients of p. This result implied a nearly polynomial-time *query* algorithm for PAC learning DNF formulas with respect to the uniform distribution.

Mansour conjectured (Mansour, 1994) that the above bound could be improved to $t^{O(\log 1/\epsilon)}$. Such an improvement would imply a polynomial-time query algorithm for learning DNF formulas with respect to the uniform distribution (to within any constant accuracy), and learning DNF formulas in this model was a major open problem at that time.

In a celebrated work from 1994, Jeff Jackson proved that DNF formulas were learnable in polynomial time (with queries, with respect to the uniform distribution) *without* proving the Mansour conjecture. His

"Harmonic Sieve" algorithm (Jackson, 1997) used boosting in combination with some weak approximation properties of polynomials. As such, for several years, Mansour's conjecture remained open and attracted considerable interest, but its resolution did not imply any new results in learning theory.

In 2008, Gopalan et al. (2008b) proved that a positive resolution to the Mansour conjecture also implies an efficient query algorithm for *agnostically* learning DNF formulas (to within any constant error parameter). The agnostic model of learning is a challenging learning scenario that requires the learner to succeed in the presence of adversarial noise. Roughly, Gopalan et al. (2008b) showed that if a class of Boolean functions $\mathcal{C}$ can be ϵ-approximated by polynomials of sparsity s, then there is a query algorithm for agnostically learning $\mathcal{C}$ in time $\mathrm{poly}(s, 1/\epsilon)$ (since decision trees are approximated by sparse polynomials, they obtained the first query algorithm for agnostically learning decision trees with respect to the uniform distribution on $\{0,1\}^n$). Whether DNF formulas can be agnostically learned (with queries, with respect to the uniform distribution) still remains a difficult open problem (Gopalan et al., 2008a).

1.1 Our Results

We prove that the Mansour conjecture is true for several well-studied subclasses of DNF formulas. As far as we know, prior to this work, the Mansour conjecture was not known to be true for any interesting class of DNF formulas.

Our first result shows that the Mansour conjecture is true for the class of randomly chosen DNF formulas:

Theorem 1 *Let $f : \{0,1\}^n \to \{0,1\}$ be a DNF formula with $t = n^{O(1)}$ terms where each term is chosen independently from the set of all terms of length $\lfloor \log t \rfloor$. Then with probability $1 - n^{-\Omega(1)}$ (over the choice of the DNF formula), there exists a polynomial p with sparsity $t^{O(\log 1/\epsilon)}$ such that $\mathbf{E}[(p(x) - f(x))^2] \leq \epsilon$.*

For $t = n^{\Theta(1)}$, the conclusion of Theorem 1 holds with probability at least $1 - n^{-\Omega(\log t)}$. Our second result is that the Mansour conjecture is true for the class of read-k DNF formulas:

Theorem 2 *Let $f : \{0,1\}^n \to \{0,1\}$ be a DNF formula with t terms where each literal appears at most k times. Then there exists a polynomial p with sparsity $t^{O(16^k \log 1/\epsilon)}$ such that $\mathbf{E}[(p(x) - f(x))^2] \leq \epsilon$.*

Even for the case $k = 1$, Mansour's conjecture was not known to be true. Mansour (1995) proves that any polynomial that approximates read-once DNF formulas to ϵ accuracy must have *degree* at least $d = \Omega(\log t \log(1/\epsilon)/\log\log(1/\epsilon))$. He further shows that a "low-degree" strategy of selecting all of a DNF formula's Fourier coefficients of monomials up to degree d results in a polynomial p with sparsity $t^{O(\log\log t \log 1/\epsilon)}$ for $1/\epsilon = \mathrm{poly}\, n$. It is not clear, however, how to improve this to the desired $t^{O(\log 1/\epsilon)}$ bound.

As mentioned earlier, by applying the result of Gopalan et al. (2008b), we obtain the first polynomial-time query algorithms for agnostically learning the above classes of DNF formulas to within any constant accuracy parameter. We consider this an important step towards agnostically learning all DNF formulas.

Corollary 3 *Let $\mathcal{C}$ be the class of DNF formulas with $t = n^{O(1)}$ terms where each term is randomly chosen from the set of all terms of length $\lfloor \log t \rfloor$. Then there is a query-algorithm for agnostically learning $\mathcal{C}$ with respect to the uniform distribution on $\{0,1\}^n$ to accuracy ϵ in time $\mathrm{poly}(n) \cdot t^{O(\log 1/\epsilon)}$ with probability $1 - n^{-\Omega(1)}$ (over the choice of the DNF formula).*

We define the notion of agnostic learning with respect to randomly chosen concept classes in Section 2. For $t = n^{\Theta(1)}$, Corollary 3 holds for a $1 - n^{-\Omega(\log t)}$ fraction of randomly chosen DNF formulas. We also obtain a corresponding agnostic learning algorithm for read-k DNF formulas:

Corollary 4 *Let $\mathcal{C}$ be the class of read-k DNF formulas with t terms. Then there is a query-algorithm for agnostically learning $\mathcal{C}$ with respect to the uniform distribution on $\{0,1\}^n$ to accuracy ϵ in time $\mathrm{poly}(n) \cdot t^{O(16^k \log 1/\epsilon)}$.*

Our sparse polynomial approximators can also be used in conjunction with recent work due to De et al. (2009) to show that for randomly chosen polynomial-size DNF formulas or read-k DNF formulas f, a $t^{-O(\log 1/\epsilon)}$-biased distribution fools f (for $k = O(1)$):

Theorem 5 *Let f be a randomly chosen polynomial-size DNF formula or a read-k DNF formula. Then (with probability $1 - n^{-\Omega(1)}$ for random DNF formulas) there exists a pseudorandom generator G such that*

$$\left| \Pr_{x \in \{0,1\}^s}[f(G(x)) = 1] - \Pr_{z \in \{0,1\}^n}[f(z) = 1] \right| \leq \epsilon$$

with $s = O(\log n + \log t \cdot \log(1/\epsilon))$.

Previously it was only known that these types of biased distributions fool read-once DNF formulas (De et al., 2009).

1.2 Related Work

As mentioned earlier, Mansour, using the random restriction machinery of Håstad (1986) and Linial et al. (1993) had shown that for any DNF formula f, there exists a polynomial of sparsity $t^{O(\log\log t \log 1/\epsilon)}$ that approximates f.

The subclasses of DNF formulas that we show are agnostically learnable have been well-studied in the PAC model of learning. Monotone read-k DNF formulas were shown to be PAC-learnable with respect to the uniform distribution by Hancock and Mansour (1991), and random DNF formulas were recently shown to be learnable on average with respect to the uniform distribution in the following sequence of work (Jackson & Servedio, 2005; Jackson et al., 2008; Sellie, 2008; Sellie, 2009).

Recently (and independently) De et al. (2009) proved that for any read-once DNF formula f, there exists an approximating polynomial p of sparsity $t^{O(\log 1/\epsilon)}$. More specifically, De et al. (2009) showed that for any class of functions $\mathcal{C}$ fooled by δ-biased sets, there exist sparse, sandwiching polynomials for $\mathcal{C}$ where the sparsity depends on δ. Since they show that $t^{-O(\log 1/\epsilon)}$-biased sets fool read-once DNF formulas, the existence of a sparse approximator for the read-once case is implicit in their work.

1.3 Our Approach

As stated above, our proof does not analyze the Fourier coefficients of DNF formulas, and our approach is considerably simpler than the random-restriction method taken by Mansour (we consider the lack of Fourier analysis a feature of the proof, given that all previous work on this problem has been Fourier-based). Instead, we use polynomial interpolation.

A Basic Example. Consider a DNF formula $f = T_1 \vee \cdots \vee T_t$ where each T_i is on a disjoint set of exactly $\log t$ variables (assume t is a power of 2). The probability that each term is satisfied is $1/t$, and the expected number of satisfied terms is one. Further, since the terms are disjoint, with high probability over the choice of the random input, only a few—say d—terms will be satisfied. As such, we construct a univariate polynomial p with $p(0) = 0$ and $p(i) = 1$ for $1 \leq i \leq d$. Then $p(T_1 + \cdots + T_t)$ will be exactly equal to f as long as at most d terms are satisfied. A careful calculation shows that the inputs where p is incorrect will not contribute too much to $\mathbf{E}[(f-p)^2]$, as there are few of them. Setting parameters appropriately yields a polynomial p that is both sparse and an ϵ-approximator of f.

Random and read-once DNF formulas. More generally, we adopt the following strategy: given a DNF formula f (randomly chosen or read-once) either (1) with sufficiently high probability a random input does not satisfy too many terms of f or (2) f is highly biased. In the former case we can use polynomial interpolation to construct a sparse approximator and in the latter case we can simply use the constant 0 or 1 function.

The probability calculations are a bit delicate, as we must ensure that the probability of many terms being satisfied decays faster than the growth rate of our polynomial approximators. For the case of random DNF formulas, we make use of some recent work due to Jackson et al. (2008) on learning random monotone DNF formulas.

Read-k DNF formulas. Read-k DNF formulas do not fit into the above dichotomy so we do not use the sum $T_1 + \cdots + T_t$ inside the univariate polynomial. Instead, we use a sum of *formulas* (rather than terms) based on a construction from (Razborov, 2008). We modify Razborov's construction to exploit the fact that terms in a read-k DNF formula do not share variables with many other terms. Our analysis shows that we can then employ the previous strategy: either (1) with sufficiently high probability a random input does not satisfy too many formulas in the sum or (2) f is highly biased.

2 Preliminaries

In this paper, we will primarily be concerned with Boolean functions $f : \{0,1\}^n \to \{0,1\}$. Let $x_1, \ldots, x_n$ be Boolean variables. A *literal* is either a variable x_i of its negation $\bar{x}_i$, and a *term* is a conjunction of literals. Any Boolean function can be expressed as a disjunction of terms, and such a formula is said to be a *disjunctive normal form* (or DNF) formula. A read-k DNF formula is a DNF formula in which the maximum number of occurrences of each variable is bounded by k. A Boolean function is monotone if changing the value of an input bit from 0 to 1 never causes the value of f to change from 1 to 0. The following consequence (Kleitman, 1966; Alon & Spencer, 2000) of the Four Functions Theorem will be useful in our study of monotone functions.

Theorem 6 *Let e, f, $\neg g$, and $\neg h$ be monotone Boolean functions over $\{0,1\}^n$. Then for any product distribution $\mathcal{D}$ over $\{0,1\}^n$, $\Pr_{\mathcal{D}}[e \wedge f] \geq \Pr_{\mathcal{D}}[e]\Pr_{\mathcal{D}}[f]$, $\Pr_{\mathcal{D}}[g \wedge h] \geq \Pr_{\mathcal{D}}[g]\Pr_{\mathcal{D}}[h]$, and $\Pr_{\mathcal{D}}[f \wedge g] \leq \Pr_{\mathcal{D}}[f]\Pr_{\mathcal{D}}[g]$.*

2.1 Sparse Polynomials

Every function $f : \{0,1\}^n \to \mathbb{R}$ can be expressed by its Fourier expansion: $f(x) = \sum_S \hat{f}(S)\chi_S(x)$ where $\chi_S(x) = \prod_{i \in S}(-1)^{x_i}$ for $S \subseteq [n]$, and $\hat{f}(S) = \mathbf{E}[f \cdot \chi_S]$. The Fourier expansion of f can be thought of as the unique polynomial representation of f over $\{+1,-1\}^n$ under the map $x_i \mapsto 1 - 2x_i$.

Mansour conjectured that polynomial-size DNF formulas could be approximated by *sparse* polynomials over $\{+1,-1\}^n$. We say a polynomial $p : \{+1,-1\}^n \to \mathbb{R}$ has sparsity s if it has at most s non-zero coefficients. We state Mansour's conjecture as originally posed in (Mansour, 1994), which uses the convention of representing FALSE by $+1$ and TRUE by -1.

Conjecture 7 (Mansour, 1994) *Let $f : \{+1,-1\}^n \to \{+1,-1\}$ be any function computable by a t-term DNF formula. Then there exists a polynomial $p : \{+1,-1\}^n \to \mathbb{R}$ with $t^{O(\log 1/\epsilon)}$ terms such that $\mathbf{E}[(f - p)^2] \leq \epsilon$.*

We will prove the conjecture to be true for various subclasses of polynomial-size DNF formulas. In our setting, Boolean functions will output 0 for FALSE and 1 for TRUE. However, we can easily change the range by setting $f^{\pm} := 1 - 2 \cdot f$. Changing the range to $\{+1,-1\}$ changes the accuracy of the approximation by at most a factor of 4: $\mathbf{E}[((1-2f) - (1-2p))^2] = 4\,\mathbf{E}[(f-p)^2]$, and it increases the sparsity by at most 1.

Given a Boolean function f, we construct a sparse approximating polynomial over $\{+1,-1\}^n$ by giving an approximating polynomial $p : \{0,1\}^n \to \mathbb{R}$ with real coefficients that has small spectral norm. The rest of the section gives us some tools to construct such polynomials and explains why doing so yields sparse approximators.

Definition 8 *The* Fourier ℓ_1-norm *(also called the* spectral norm*) of a function $p : \{0,1\}^n \to \mathbb{R}$ is defined to be $\|p\|_1 := \sum_S |\hat{p}(S)|$. We will also use the following minor variant, $\|p\|_1^{\neq \emptyset} := \sum_{S \neq \emptyset} |\hat{p}(S)|$.*

The following two facts about the spectral norm of functions will allow us to construct polynomials over $\{0,1\}^n$ naturally from DNF formulas.

Fact 9 *Let $p : \{0,1\}^m \to \mathbb{R}$ be a polynomial with coefficients $p_S \in \mathbb{R}$ for $S \subseteq [m]$, and $q_1, \ldots, q_m : \{0,1\}^n \to \{0,1\}$ be arbitrary Boolean functions. Then $p(q_1, \ldots, q_m) = \sum_S p_S \prod_{i \in S} q_i$ is a polynomial over $\{0,1\}^n$ with spectral norm at most*

$$\sum_{S \subseteq [m]} |p_S| \prod_{i \in S} ||q_i||_1.$$

Proof: The fact follows by observing that for any $p, q : \{0,1\}^n \to \mathbb{R}$, we have $||p+q||_1 \leq ||p||_1 + ||q||_1$ and $||pq||_1 \leq ||p||_1 ||q||_1$. ■

Fact 10 *Let $T : \{0,1\}^n \to \{0,1\}$ be an AND of a subset of its literals. Then $||T||_1 = 1$.*

Finally, using the fact below, we show why approximating polynomials with small spectral norm give sparse approximating polynomials.

Fact 11 (Kushilevitz & Mansour, 1993) *Given any function $f : \{0,1\}^n \to \mathbb{R}$ and $\epsilon > 0$, let $\mathcal{S} = \{S \subseteq [n] : |\hat{f}(S)| \geq \epsilon/\|f\|_1\}$, and $g(x) = \sum_{S \in \mathcal{S}} \hat{f}(S)\chi_s(x)$. Then $\mathbf{E}[(f-g)^2] \leq \epsilon$, and $|\mathcal{S}| \leq \|f\|_1^2/\epsilon$.*

Now, given functions $f, p : \{0,1\}^n \to \mathbb{R}$ such that $E[(f-p)^2] \leq \epsilon$, we can construct a 4ϵ-approximator for f with sparsity $||p||_1^2/\epsilon$ by defining $p'(x) = \sum_{S \in \mathcal{S}} \hat{p}(S)\chi_S(x)$ as in Fact 11. Clearly p' has sparsity $||p||_1^2/\epsilon$, and

$$E[(f-p')^2] = E[(f - p + p - p')^2] \leq E[2((f-p)^2 + (p-p')^2)] \leq 4\epsilon,$$

where the first inequality follows from the inequality $(a+b)^2 \leq 2(a^2+b^2)$ for any reals a and b.

2.2 Agnostic learning

We first describe the traditional framework for agnostically learning concept classes with respect to the uniform distribution and then give a slightly modified definition for an "average-case" version of agnostic learning where the unknown concept (in this case a DNF formula) is randomly chosen.

Definition 12 (Standard agnostic model) *Let $f : \{+1,-1\}^n \to \{+1,-1\}$ be an arbitrary function, and let $\mathcal{D}$ be the uniform distribution on $\{+1,-1\}^n$. Define*

$$\mathsf{opt} = \min_{c \in \mathcal{C}} \Pr_{x \sim \mathcal{D}}[c(x) \neq f(x)].$$

That is, opt *is the error of the best fitting concept in $\mathcal{C}$ with respect to $\mathcal{D}$. We say that an algorithm A agnostically learns $\mathcal{C}$ with respect to $\mathcal{D}$ if the following holds for any f: if A is given black-box access to f then with high probability A outputs a hypothesis h such that $\Pr_{x \sim \mathcal{D}}[h(x) \neq f(x)] \leq \mathsf{opt} + \epsilon$.*

The intuition behind the above definition is that a learner—given access to a concept $c \in \mathcal{C}$ where an η fraction of c's inputs have been adversarially corrupted—should still be able to output a hypothesis with accuracy $\eta + \epsilon$ (achieving error better than η may not be possible, as the adversary could embed a completely random function on an η fraction of c's inputs). Here η plays the role of opt.

This motivates the following definition for agnostically learning a randomly chosen concept from some class $\mathcal{C}$:

Definition 13 (Agnostically learning random concepts) *Let $\mathcal{C}$ be a concept class and choose c randomly from $\mathcal{C}$ (the distribution over $\mathcal{C}$ will be clear from the context). We say that an algorithm A agnostically learns random concepts from $\mathcal{C}$ if with probability at least $1 - \delta$ over the choice of c the following holds: if the learner is given black-box access to some fixed function c' and $\Pr_{x \in \{+1,-1\}^n}[c(x) \neq c'(x)] \leq \eta$, then A outputs a hypothesis h such that $\Pr_{x \in \{+1,-1\}^n}[h(x) \neq c'(x)] \leq \eta + \epsilon$.*

We are unaware of any prior work defining an agnostic framework for learning randomly chosen concepts.

The main result we use to connect the approximation of DNF formulas by sparse polynomials with agnostic learning is due to Gopalan et al. (2008b):

Theorem 14 (Gopalan et al., 2008b) *Let $\mathcal{C}$ be a concept class such that for every $c \in \mathcal{C}$ there exists a polynomial p such that $\|p\|_1 \leq s$ and $\mathbf{E}_{x \in \{+1,-1\}^n}[|p(x) - c(x)|^2] \leq \epsilon^2/2$. Then there exists an algorithm B such that the following holds: given black-box access to any Boolean function $f : \{+1,-1\}^n \to \{+1,-1\}$, B runs in time $\mathrm{poly}(n, s, 1/\epsilon)$ and outputs a hypothesis $h : \{+1,-1\}^n \to \{+1,-1\}$ with*

$$\Pr_{x \in \{+1,-1\}^n}[h(x) \neq f(x)] \leq \mathsf{opt} + \epsilon.$$

3 Approximating DNFs using univariate polynomial interpolation

Let $f = T_1 \vee T_2 \vee \cdots \vee T_t$ be any DNF formula. We say $T_i(x) = 1$ if x satisfies the term T_i, and 0 otherwise. Let $y_f : \{0,1\}^n \to \{0,\ldots,t\}$ be the function that outputs the number of terms of f satisfied by x, *i.e.,* $y_f(x) = T_1(x) + T_2(x) + \cdots + T_t(x)$.

Our constructions will use the following univariate polynomial P_d to interpolate the values of f on inputs $\{x : y_f(x) \leq d\}$.

Fact 15 *Let*

$$P_d(y) := (-1)^{d+1} \frac{(y-1)(y-2)\cdots(y-d)}{d!} + 1. \tag{1}$$

Then, (1) the polynomial P_d is a degree-d polynomial in y; (2) $P_d(0) = 0$, $P_d(y) = 1$ for $y \in [d]$, and for $y \in [t] \setminus [d]$, $P_d(y) = -\binom{y-1}{d} + 1 \leq 0$ if d is even and $P_d(y) = \binom{y-1}{d} + 1 > 1$ if d is odd; and (3) the sum of the magnitudes of P_d's coefficients is d.

Proof: Properties (1) and (2) can be easily verified by inspection. Expanding the falling factorial, we get that $(y-1)(y-2)\cdots(y-d) = \sum_{j=0}^{d} (-1)^{d-j} \left[{d+1 \atop j+1}\right] y^j$, where $\left[{a \atop b}\right]$ denotes a Stirling number of the first kind. The Stirling numbers of the first kind count the number of permutations of a elements with b disjoint cycles. Therefore, $\sum_{j=0}^{d} \left[{d+1 \atop j+1}\right] = (d+1)!$ (Graham et al., 1994). The constant coefficient of P_d is 0 by Property (2), thus the sum of the absolute values of the other coefficients is $((d+1)! - d!)/d! = d$. ∎

For any t-term DNF formula f, we can construct a polynomial $p_{f,d} : \{0,1\}^n \to \mathbb{R}$ defined as $p_{f,d} := P_d \circ y_f$. A simple calculation, given below, shows that the ℓ_1-norm of $p_{f,d}$ is polynomial in t and exponential in d.

Lemma 16 *Let f be a t-term DNF formula, then $\|p_{f,d}\|_1 \leq t^{O(d)}$.*

Proof: By Fact 15, P_d is a degree-d univariate polynomial with d non-zero coefficients of magnitude at most d. We can view the polynomial $p_{f,d}$ as the polynomial $P'_d(T_1, \ldots, T_t) := P_d(T_1 + \cdots + T_t)$ over variables $T_i \in \{0, 1\}$. Expanding out P'_d gives us at most dt^d monomials with coefficients of magnitude at most d. Now each monomial of P'_d is a product of T_i's, so applying Facts 10 and 9 we have that $\|p_{f,d}\|_1 \leq t^{O(d)}$. ■

The next two sections will show that the polynomial $p_{f,d}$ (for $d = \Theta(\log 1/\epsilon)$) is in fact a good approximation for random DNF formulas and (with a slight modification) for read-k DNF formulas. As a warm-up, we will show the simple case of read-once DNF formulas.

3.1 A Simple Case: Read-Once DNF Formulas

For read-once DNF formulas, the probability that a term is satisfied is independent of whether or not any of the other terms are satisfied, and thus f is unlikely to have many terms satisfied simultaneously.

Lemma 17 *Let $f = T_1 \vee, \cdots, \vee T_t$ be a read-once DNF formula of size t such that $\Pr[f] < 1 - \epsilon$. Then the probability over the uniform distribution on $\{0, 1\}^n$ that some set of $j > e \ln 1/\epsilon$ terms is satisfied is at most $\left(\frac{e \ln 1/\epsilon}{j}\right)^j$.*

Proof: For any assignment x to the variables of f, let $y_f(x)$ be the number terms satisfied in f. By linearity of expectation, we have that $\mathbf{E}_x[y_f(x)] = \sum_{i=1}^t \Pr[T_i = 1]$. Note that $\Pr[\neg f] = \prod_{i=1}^t (1 - \Pr[T_i])$, which is maximized when each $\Pr[T_i] = \mathbf{E}[y_f]/t$, hence $\Pr[\neg f] \leq (1 - \mathbf{E}[y_f]/t)^t \leq e^{-\mathbf{E}[y_f]}$. Thus we may assume that $\mathbf{E}[y_f] \leq \ln 1/\epsilon$, otherwise $\Pr[f] \geq 1 - \epsilon$.

Assuming $\mathbf{E}[y_f] \leq \ln 1/\epsilon$, we now bound the probability that some set of $j > e \ln 1/\epsilon$ terms of f is satisfied. Since all the terms are disjoint, this probability is $\sum_{S \subseteq [t], |S| = j} \prod_{i \in S} \Pr[T_i]$, and the arithmetic-geometric mean inequality gives that this is maximized when every $\Pr[T_i] = \mathbf{E}[y_f]/t$. Then the probability of satisfying some set of j terms is at most:

$$\binom{t}{j} \left(\frac{\ln 1/\epsilon}{t}\right)^j \leq \left(\frac{et}{j}\right)^j \left(\frac{\ln 1/\epsilon}{t}\right)^j = \left(\frac{e \ln 1/\epsilon}{j}\right)^j,$$

which concludes the proof of the lemma. ■

The following lemma shows that we can set d to be fairly small, $\Theta(\log 1/\epsilon)$, and the polynomial $p_{f,d}$ will be a good approximation for any DNF formula f, as long as f is unlikely to have many terms satisfied simultaneously.

Lemma 18 *Let f be any t-term DNF formula, and let $d = \lceil 4e^3 \ln 1/\epsilon \rceil$. If*

$$\Pr[y_f(x) = j] \leq \left(\frac{e \ln 1/\epsilon}{j}\right)^j$$

for every $d \leq j \leq t$, then the polynomial $p_{f,d}$ satisfies $\mathbf{E}[(f - p_{f,d})^2] \leq \epsilon$.

Proof: We condition on the values of $y_f(x)$, controlling the magnitude of $p_{f,d}$ by the unlikelihood of y_f being large. By Fact 15, $p_{f,d}(x)$ will output 0 if x does not satisfy f, $p_{f,d}(x)$ will output 1 if $y_f(x) \in [d]$, and $|p_{f,d}(x)| < \binom{y_f}{d}$ for $y_f(x) \in [t] \setminus [d]$. Hence:

$$\begin{aligned}
\|f - p_{f,d}\|^2 &< \sum_{j=d+1}^{t} \binom{j}{d}^2 \left(\frac{e \ln 1/\epsilon}{j}\right)^j \\
&< \sum_{j=d+1}^{t} 2^{2j} \left(\frac{e \ln 1/\epsilon}{4e^3 \ln 1/\epsilon}\right)^j \\
&< \epsilon \sum_{j=d+1}^{t} \frac{1}{e^j} < \epsilon.
\end{aligned}$$

■

Combining Lemmas 16, 17, and 18 gives us Mansour's conjecture for read-once DNF formulas.

Theorem 19 *Let f be any read-once DNF formula with t terms. Then there is a polynomial $p_{f,d}$ with $\|p_{f,d}\|_1 \leq t^{O(\log 1/\epsilon)}$ and $\mathbf{E}[(f - p_{f,d})^2] \leq \epsilon$ for all $\epsilon > 0$.*

4 Mansour's Conjecture for Random DNF Formulas

In this section, we establish various properties of random DNF formulas and use these properties to show that for almost all f, Mansour's conjecture holds. Roughly speaking, we will show that a random DNF formula behaves like a read-once DNF formula, in that any "large" set of terms is unlikely to be satisfied by a random assignment. This notion is formalized in Lemma 22. For such DNF formulas, we may use the construction from Section 3 to obtain a good approximating polynomial for f with small spectral norm (Theorem 24).

Throughout the rest of this section, we assume that $t(n) = n^{O(1)}$. For brevity we write t for $t(n)$. Let $\mathcal{D}_n^t$ be the probability distribution over t-term DNF formulas induced by the following process: each term is independently and uniformly chosen at random from all $t\binom{n}{\log t}$ possible terms of size exactly $\log t$ over $\{x_1, \ldots, x_n\}$. For convenience, we assume that $\log t$ is an integer throughout our discussion, although the general case is easily handled by taking terms of length $\lfloor \log t \rfloor$. If the terms are not of size $\Theta(\log n)$, then the DNF will be biased, and thus be easy to learn. We refer the reader to Jackson and Servedio (2005) for a full discussion of the model.

If t grows very slowly relative to n, say $t = n^{o(1)}$, then with high probability $(1 - n^{\Omega(1)})$ a random f drawn from $\mathcal{D}_n^t$ will be a read-once DNF formula, in which case the results of Section 3.1 hold. Therefore, throughout the rest of this section we will assume that t is in fact $n^{\Theta(1)}$.

To prove Lemma 22, we require two lemmas, which are inspired by the results of (Jackson & Servedio, 2005) and (Jackson et al., 2008). Lemma 20 shows that with high probability the terms of a random DNF formula are close to being disjoint, and thus cover close to $j \log t$ variables.

Lemma 20 *With probability at least $1 - t^j e^{j \log t} (j \log t)^{\log t} / n^{\log t}$ over the random draw of f from $\mathcal{D}_n^t$, at least $j \log t - (\log t)/4$ variables occur in every set of j distinct terms of f. The failure probability is at most $1/n^{\Omega(\log t)}$ for any $j < c \log n$, for some constant c.*

Proof: Let $k := \log t$. Fix a set of j terms, and let $v \leq jk$ be the number of distinct variables (negated or not) that occur in these terms. We will bound the probability that $v > w := jk - k/4$. Consider any particular fixed set of w variables. The probability that none of the j terms include any variable outside of the w variables is precisely $\left(\binom{w}{k}/\binom{n}{k}\right)^j$. Thus, the probability that $v \leq w$ is by the union bound:

$$\binom{n}{w}\left(\frac{\binom{w}{k}}{\binom{n}{k}}\right)^j < \left(\frac{en}{w}\right)^w \left(\frac{w}{n}\right)^{jk} = \frac{e^{jk-k/4}(jk-k/4)^{k/4}}{n^{k/4}} < \frac{e^{jk}(jk)^{k/4}}{n^{k/4}}.$$

Taking a union bound over all (at most t^j) sets, we have that with the correct probability every set of j terms contains at least w distinct variables. ■

We will use the method of bounded differences (a.k.a., McDiarmid's inequality) to prove Lemma 22.

Proposition 21 (McDiarmid's inequality) *Let $X_1, \ldots, X_m$ be independent random variables taking values in a set $\mathcal{X}$, and let $f : \mathcal{X}^m \to \mathbb{R}$ be such that for all $i \in [m]$, $|f(a) - f(a')| \leq d_i$, whenever $a, a' \in \mathcal{X}^m$ differ in just the ith coordinate. Then for all $\tau > 0$,*

$$\Pr[f > \mathbf{E} f + \tau] \leq \exp\left(-\frac{2\tau^2}{\sum_i d_i^2}\right) \text{ and } \Pr[f < \mathbf{E} f - \tau] \leq \exp\left(-\frac{2\tau^2}{\sum_i d_i^2}\right).$$

The following lemma shows that with high probability over the choice of random DNF formula, the probability that exactly j terms are satisfied is close to that for the "tribes" function: $\binom{t}{j} t^{-j}(1-1/t)^{t-j}$.

Lemma 22 *There exists a constant c such that for any $j < c \log n$, with probability at least $1 - 1/n^{\Omega(\log t)}$ over the random draw of f from $\mathcal{D}_n^t$, the probability over the uniform distribution on $\{0,1\}^n$ that an input satisfies exactly j distinct terms of f is at most $2\binom{t}{j} t^{-j}(1-1/t)^{t-j}$.*

Proof: Let $f = T_1 \vee \cdots \vee T_t$, and let $\beta := t^{-j}(1-1/t)^{t-j}$. Fix any $J \subset [t]$ of size j, and let U_J be the probability over $x \in \{0,1\}^n$ that the terms T_i for $i \in J$ are satisfied and no other terms are satisfied. We will show that $U_J < 2\beta$ with high probability; a union bound over all possible sets J of size j in $[t]$ gives that $U_J \leq 2\beta$ for every J with high probability. Finally, a union bound over all $\binom{t}{j}$ possible sets of j terms (where the probability is taken over x) proves the lemma.

Without loss of generality, we may assume that $J = [j]$. For any fixed x, we have:

$$\Pr_{f \in \mathcal{D}_n^t}[x \text{ satisfies exactly the terms in } J] = \beta,$$

and thus by linearity of expectation, we have $\mathbf{E}_{f\in\mathcal{D}_n^t}[U_J] = \beta$. Now we show that with high probability that the deviation of U_J from its expected value is low.

Applying Lemma 20, we may assume that the terms $T_1, \cdots, T_j$ contain at least $j\log t - (\log t)/4$ many variables, and that $J \cup T_i$ for all $i = j+1, \cdots, t$ includes at least $(j+1)\log t - (\log t)/4$ many unique variables, while increasing the failure probability by only $1/n^{\Omega(\log t)}$. Note that conditioning on this event can change the value of U_J by at most $1/n^{\Omega(\log t)} < \frac{1}{2}\beta$, so under this conditioning we have $\mathbf{E}[P_j] \geq \frac{1}{2}\beta$. Conditioning on this event, fix the terms $T_1, \cdots, T_j$. Then the terms $T_{j+1}, \cdots, T_t$ are chosen uniformly and independently from the set of all terms T of length $\log t$ such that the union of the variables in J and T includes at least $(j+1)\log t - (\log t)/4$ unique variables. Call this set $\mathcal{X}$.

We now use McDiarmid's inequality where the random variables are the terms $T_{j+1}, \ldots, T_t$ randomly selected from $\mathcal{X}$, letting $g(T_{j+1}, \cdots, T_t) = U_J$ and $g(T_{j+1}, \cdots, T_{i-1}, T_i', T_{i+1}, \cdots, T_t) = U_J'$ for all $i = j+1, \ldots, t$. We claim that:

$$|U_J - U_J'| \leq d_i := \frac{t^{1/4}}{t^{j+1}}.$$

This is because U_J' can only be larger than U_J by assignments which satisfy $T_1, \cdots, T_J$ and T_i. Similarly, U_J' can only be smaller than U_J by assignments which satisfy $T_1, \cdots, T_J$ and T_i'. Since T_i and T_i' come from $\mathcal{X}$, we know that at least $(j+1)t - (\log t)/4$ variables must be satisfied.

Thus we may apply McDiarmid's inequality with $\tau = \frac{3}{2}\beta$, which gives that $\Pr_f[U_J > 2\beta]$ is at most

$$\exp\left(\frac{-2\frac{9}{4}\beta^2}{t^{3/2}/t^{2j+2}}\right) \leq \exp\left(\frac{-9\sqrt{t}(1-1/t)^{2(t-j)}}{2}\right).$$

Combining the failure probabilities over all the $\binom{t}{j}$ possible sets, we get that with probability at least

$$\binom{t}{j}\left(\frac{1}{n^{\Omega(\log t)}} + e^{-9\sqrt{t}(1-1/t)^{2(t-j)}/2}\right) = \frac{1}{n^{\Omega(\log t)}},$$

over the random draw of f from $\mathcal{D}_n^t$, U_J for all $J \subseteq [t]$ of size j is at most 2β. Thus, the probability that a random input satisfies exactly some j distinct terms of f is at most $2\binom{t}{j}\beta$. ■

Using these properties of random DNF formulas we can now show a lemma analogous to Lemma 18 for random DNF formulas.

Lemma 23 *Let f be any DNF formula with $t = n^{O(1)}$ terms, and let $\epsilon > 0$ which satisfies $1/\epsilon = o(\log\log n)$. Then set $d = \lceil 4e^3 \ln 1/\epsilon \rceil$ and $\ell = c\log n$, where c is the constant in Lemma 22. If*

$$\Pr[y_f(x) = j] \leq \left(\frac{e\ln 1/\epsilon}{j}\right)^j$$

for every $d \leq j \leq \ell$, then the polynomial $p_{f,d}$ satisfies $\mathbf{E}[(f - p_{f,d})^2] \leq \epsilon$.

Proof: We condition on the values of $y_f(x)$, controlling the magnitude of $p_{f,d}$ by the unlikelihood of y_f being large. By Fact 15, $p_{f,d}(x)$ will output 0 if x does not satisfy f, $p_{f,d}(x)$ will output 1 if $y_f(x) \in [d]$, and $|p_{f,d}(x)| < \binom{y_f}{d}$ for $y_f(x) \in [t] \setminus [d]$. Hence:

$$\begin{aligned}
\|f - p_{f,d}\|^2 &< \sum_{j=d+1}^{\ell-1}\binom{j}{d}^2\left(\frac{e\ln 1/\epsilon}{j}\right)^j + \binom{t}{d}^2 \cdot \Pr[y_f \geq \ell] \\
&< \sum_{j=d+1}^{\ell-1} 2^{2j}\left(\frac{e\ln 1/\epsilon}{4e^3\ln 1/\epsilon}\right)^j + n^{-\Omega(\log\log n)} \\
&< \epsilon\sum_{j=d+1}^{\ell-1}\frac{1}{e^j} + n^{-\Omega(\log\log n)} < \epsilon.
\end{aligned}$$

■

We can now show that Mansour's conjecture (Mansour, 1994) is true with high probability over the choice of f from $\mathcal{D}_n^t$.

Theorem 24 *Let $f : \{0,1\}^n \to \{0,1\}$ be a $t = n^{\Theta(1)}$-term DNF formula where each term is chosen independently from the set of all terms of length $\log t$. Then with probability at least $1 - n^{-\Omega(\log t)}$ over the choice of f, there exists a polynomial p with $\|p\|_1 \leq t^{O(\log 1/\epsilon)}$ such that $\mathbf{E}[(p(x) - f(x))^2] \leq \epsilon$.*

Proof: Let $d := \lceil 4e^3 \ln(1/\epsilon) \rceil$ and $p_{f,d}$ be as defined in Section 3. Lemma 16 tells us that $\|p_{f,d}\|_1 \leq t^{O(\log 1/\epsilon)}$. We show that with probability at least $1 - n^{-\Omega(\log t)}$ over the random draw of f from $\mathcal{D}_n^t$, $p_{f,d}$ will be a good approximator for f. This follows by Lemma 22; with probability at least $1 - (c\log(n) - d - 1)/n^{\Omega(\log t)} = 1 - n^{-\Omega(\log t)}$, we have $\Pr[y = j]$ for all $d < j \leq c\log(n)$. Thus for such f Lemma 18 tells us that $\mathbf{E}[(f - p_{f,d})^2] \leq \epsilon$. ■

5 Mansour's Conjecture for Read-k DNF Formulas

In this section, we give an ϵ-approximating polynomial for any read-k DNF formula and show that its spectral norm is at most $t^{O(2^{4k}\log 1/\epsilon)}$. This implies that Mansour's conjecture holds for all read-k DNF formulas where k is any constant.

Read-k DNF formulas may not satisfy the conditions of Lemma 18, so we must change our approach. Instead of using $\sum_{i=1}^t T_i$ inside our univariate polynomial, we use a different sum, which is based on a construction from (Razborov, 2008) for representing any DNF formula. We modify this representation to exploit the fact that for read-k DNF formulas, the variables in a term can not share variables with too many other terms. Unlike for read-once DNF formulas, it is not clear that the number of terms satisfied in a read-k DNF formula will be extremely concentrated on a small range. We show how to modify our construction so that a concentration result does hold.

Let $f = T_1 \vee \cdots \vee T_t$ be any t-term read-k DNF formula, and let $|T_i|$ denote the number of variables in the term T_i. We assume that the terms are ordered from longest to shortest, *i.e.,* $|T_j| \geq |T_i|$ for all $j \leq i$. For any term T_i of f, let ϕ_i be the DNF formula consisting of those terms (at least as large as T_i) in $T_1, \cdots, T_{i-1}$ that overlap with T_i, *i.e.,*

$$\phi_i := \bigvee_{j \in \mathcal{C}_i} T_j, \text{ for } \mathcal{C}_i = \{j < i \mid T_j \cap T_i \neq \emptyset\}.$$

We define $A_i := T_i \wedge \neg\phi_i$ and $z_f := \sum_{i=1}^t A_i$. The function $z_f : \{0,1\}^n \to \{0, \ldots, t\}$ outputs the number of disjoint terms of f satisfied by x (greedily starting from T_1). Note that if f is a read-once DNF formula, then $z_f = y_f$.

Observe that each A_i can be represented by the polynomial $T_i \cdot \prod_{j \in C_i}(1 - T_j)$ (and so z_f can be represented by a polynomial), and that $||(1 - T_j)||_1 \leq 2$ for all j. As f is a read-k DNF formula, each ϕ_i has at most $k|T_i|$ terms, and A_i has small spectral norm:

Fact 25 *Let $f = T_1 \vee \cdots \vee T_t$ be a t-term read-k DNF formula. Then each A_i has a polynomial representation, and $||A_i||_1 \leq 2^{k|T_i|}$.*

As we did in Section 3, we can construct a polynomial $q_{f,d} : \{0,1\}^n \to \mathbb{R}$ defined as $q_{f,d} := P_d \circ z_f$ for any t-term read-k DNF formula f. The following lemma shows that $q_{f,d}$ has small spectral norm.

Lemma 26 *Let f be a t-term read-k DNF formula with terms of length at most w. Then $\|q_{f,d}\|_1 \leq 2^{O(d(\log t + kw))}$.*

Proof: By Fact 15, P_d is a degree-d univariate polynomial with d terms and coefficients of magnitude at most d. We can view the polynomial $q_{f,d}$ as the polynomial $P'_d(A_1, \ldots, A_t) := P_d(A_1 + \cdots + A_t)$ over variables $A_i \in \{0,1\}$. Expanding out (but not recombining) P'_d gives us at most dt^d monomials of degree d (over variables A_i) with coefficients of magnitude at most d.

We can now apply Facts 25 and 9 to bound the spectral norm of $q_{f,d}$. Since P'_d has at most dt^d monomials each of degree d (over A_i), fnd each A_i satisfies $||A_i||_1 \leq 2^{kw}$, we have that $||q_{f,d}||_1 \leq 2^{dkw}dt^d = 2^{O(d(\log t + kw))}$. ■

We will show that Mansour's conjecture holds for read-k DNF formulas by showing that $z_f = \sum_{i=1}^t A_i$ behaves much like $y_f = \sum_{i=1}^t T_i$ would if f were a read-once DNF formula, and thus we can use our polynomial P_d (Equation 1) to approximate f.

One crucial property of our construction is that only disjoint sets of terms can contribute to z_f.

Claim 27 *Let $T_1 \vee \cdots \vee T_t$ be a t-term DNF formula. Then for any $S \subseteq [t]$, $\Pr[\wedge_{i \in S} A_i] \leq \prod_{i \in S} \Pr[T_i]$.*

Proof: If there is a pair $j, k \in S$ such that $T_j \cap T_k \neq \emptyset$ for some $j < k$, then ϕ_k contains T_j and both $T_j \wedge \neg\phi_j$ and $T_k \wedge \neg\phi_k$ cannot be satisfied simultaneously, so $\Pr[\wedge_{i \in S} A_i] = 0$. If no such pair exists, then all the terms indexed by S are disjoint. Thus,

$$\Pr[\wedge_{i \in S} A_i] \leq \Pr[\wedge_{i \in S} T_i] = \prod_{i \in S} \Pr[T_i],$$

as was to be shown. ■

The following lemma was communicated to us by Omid Etesami and James Cook (Etesami & Cook, 2010).

Lemma 28 *Let $f = T_1 \vee \ldots \vee T_t$ be a t-term read-k DNF formula, and let $f' = T'_1 \vee \ldots \vee T'_t$ be the monotone formula obtained from f by replacing all the negative literals by their positive counterparts. Then* $\Pr[f'] \leq \Pr[f]$.

Proof: For each $0 \leq i \leq n$, define $f^{(i)}$ as the DNF formula obtained from f when replacing each occurrence of $\neg x_j$ by x_j for all $1 \leq j \leq i$. In particular, $f^{(0)} = f$ and $f^{(n)} = f'$. Let $f^{(i-1)} = (g_{x_i} \wedge x_i) \vee (g_{\neg x_i} \wedge \neg x_i) \vee g_\emptyset$ where $g_{x_i} \wedge x_i$ is the OR of all terms from $f^{(i-1)}$ that have the literal x_i, $g_{\neg x_i} \wedge \neg x_i$ is the OR of all terms that have the literal $\neg x_i$, and $g_\emptyset$ is the OR of all terms that neither contain x_i nor contain $\neg x_i$. Note that $f^{(i)} = ((g_{x_i} \vee g_{\neg x_i}) \wedge x_i) \vee g_\emptyset$. Thus

$$\Pr\left[f^{(i-1)}\right] = \frac{1}{2}\Pr[g_{x_i} \wedge \neg g_\emptyset] + \frac{1}{2}\Pr[g_{\neg x_i} \wedge \neg g_\emptyset] + \Pr[g_\emptyset],$$

and

$$\Pr\left[f^{(i)}\right] = \frac{1}{2}\Pr[(g_{x_i} \vee g_{\neg x_i}) \wedge \neg g_\emptyset] + \Pr[g_\emptyset].$$

A union bound on the events $(g_{x_i} \wedge \neg g_\emptyset)$ and $(g_{\neg x_i} \wedge \neg g_\emptyset)$ tells us that $\Pr[f^{(i-1)}] \geq \Pr[f^{(i)}]$, and thus $\Pr[f^{(0)}] \geq \Pr[f^{(n)}]$. ■

As in the read-once case, we will prove that for any read-k DNF formula f, if $\sum_{i=1}^{t} \Pr[T_i]$ is large then f is biased towards one (Lemma 30). To do so we will prove this for monotone read-k DNF formulas and then use Lemma 28 to obtain the general case. Before we prove Lemma 30 we need the following claim, which tells us that for a read-k *monotone* DNF formula, the probability of satisfying A_i compared to that of satisfying T_i is only smaller by a constant (for constant k).

Claim 29 *Let $T_1 \vee \cdots \vee T_t$ be a t-term monotone read-k DNF formula. Then* $2^{-4k}\Pr[T_i] \leq \Pr[A_i]$.

Proof: Let I be the set of indices of the terms in ϕ_i. For each $T_j \in \phi_i$, let T'_j be T_j with all the variables of T_i set to 1, and let $\phi'_i = \vee_{\{j:T_j \in \phi_i\}} T'_j$. (For example, if $T_i = x_1x_2x_3$ and $T_j = x_2x_4x_5$ is a term of ϕ_i, then ϕ'_i contains the term $T'_j = x_4x_5$.) Observe that $\Pr[A_i] = \Pr[T_i \wedge \neg\phi_i] = \Pr[T_i \wedge \neg\phi'_i] = \Pr[T_i]\Pr[\neg\phi'_i]$. Thus it suffices to show that $\Pr[\neg\phi'_i] \geq 2^{-4k}$.

Let a_j be the number of variables in $T_j \cap T_i$. By the definition of ϕ_i, $1 \leq a_j \leq |T_i| - 1$, and note that $\Pr[T'_j] = 2^{a_j - |T_j|}$. Applying the Four Functions Theorem (Theorem 6), we obtain:

$$\Pr[\neg\phi'_i] \geq \prod_{j \in I} \Pr[\neg T'_j] = \prod_{j \in I}(1 - 2^{a_j - |T_j|}) \geq \prod_{j \in I}(1 - 2^{a_j - |T_i|}).$$

We partition I into two sets: $J = \{j : a_j \leq |T_i|/2\}$ and $J' = \{j : a_j > |T_i|/2\}$. (Assume that $|T_i| \geq 4$ or else we are done, because there can be at most $4k$ terms.) As ϕ_i is a read-k DNF formula, we have that $\sum_{j \in I} a_j \leq k|T_i|$, and thus $|J'| \leq 2k$, and $|J| \leq k|T_i|$.

We will lower bound the products over each set of indices separately. For those terms in J, we have that $\Pr[T'_j] \leq 2^{-|T_i|/2}$, hence

$$\prod_{j \in J}(1 - \Pr[T'_j]) \geq \prod_{j \in J}(1 - 2^{-|T_i|/2}) \geq (1 - 2^{-|T_i|/2})^{k|T_i|} \geq 2^{-2k}.$$

For those terms T_j, $j \in J'$ (which share many variables with T_i), we use the facts that each $\Pr[T'_j] \leq 1/2$ and that there are at most $2k$ such terms, so that

$$\prod_{j \in J'}(1 - \Pr[T'_j]) \geq 2^{-2k}.$$

Taking the product over the set $J \cup J'$ completes the proof of the claim. ■

Finally, we will prove that for any read-k DNF formula f, if $\sum_{i=1}^{t} \Pr[T_i]$ is large then f is biased towards one. Using Lemma 30 with Claim 27, we can prove a lemma analogous to Lemma 17 by a case analysis of $\sum_{i=1}^{t} \Pr[T_i]$; either it is large and f must be biased toward one, or it is small so z_f is usually small.

Lemma 30 *Let f be a t-term read-k DNF formula. Then,*

$$\sum_{i=1}^{t} \Pr[T_i] \leq 2^{4k} \ln\left(\frac{1}{\Pr[\neg f]}\right).$$

Proof: First, let us consider the case when f is monotone. Let ρ_i be those terms among $T_1, \ldots, T_{i-1}$ that are not present in ϕ_i. We can upper-bound $\Pr[\neg f]$ by:

$$\begin{aligned}
\Pr[\neg f] &= \prod_{i=1}^{t} (1 - \Pr[T_i \mid \neg\phi_i \wedge \neg\rho_i]) \\
&\leq \prod_{i=1}^{t} (1 - \Pr[T_i \wedge \neg\phi_i \mid \neg\rho_i]) = \prod_{i=1}^{t} (1 - \Pr[T_i \mid \neg\rho_i] \Pr[\neg\phi_i \mid T_i \wedge \neg\rho_i]) \\
&\leq \prod_{i=1}^{t} (1 - \Pr[T_i] \Pr[\neg\phi_i \mid T_i]) = \prod_{i=1}^{t} (1 - \Pr[A_i]).
\end{aligned}$$

The first inequality comes from $\Pr[A \mid B \wedge C] \geq \Pr[A \wedge B \mid C]$ for any A, B, and C. The last inequality holds because $\Pr[T_i \mid \neg\rho_i] = \Pr[T_i]$ (by the mutual independence of T_i and ρ_i) and $\Pr[\neg\phi_i \mid T_i] \leq \Pr[\neg\phi_i \mid T_i \wedge \neg\rho_i]$. The last fact may be obtained by applying the Four Functions Theorem to $\neg\phi_i$ and $\neg\rho_i$ under the product distribution induced by setting all the variables of T_i to be true.

We apply Claim 29 to obtain $\Pr[\neg f] \leq \prod_{i=1}^{t}(1 - \Pr[T_i]2^{-4k})$, and the arithmetic-geometric mean inequality shows that our upper-bound on $\Pr[\neg f]$ is maximized when all the $\Pr[T_i]$ are equal, hence:

$$\Pr[\neg f] \leq \left(1 - 2^{-4k} \frac{\sum_{i=1}^{t} \Pr[T_i]}{t}\right)^t \leq \exp\left(-2^{-4k} \sum_{i=1}^{t} \Pr[T_i]\right).$$

Solving for $\sum_{i=1}^{t} \Pr[T_i]$ yields the lemma.

Now let f be a non-monotone DNF formula, and let f' be the monotonized version of f. Then by Lemma 28 we have:

$$\sum_{i=1}^{t} \Pr[T_i] = \sum_{i=1}^{t} \Pr[T'_i] \leq 2^{4k} \ln\left(\frac{1}{\Pr[\neg f']}\right) \leq 2^{4k} \ln\left(\frac{1}{\Pr[\neg f]}\right),$$

as was to be shown. ■

Lemma 31 *Let $f = T_1 \vee \cdots \vee T_t$ be a read-k DNF formula of size t such that $\Pr[f] < 1 - \epsilon$. Then the probability over the uniform distribution on $\{0,1\}^n$ that $z_f \geq j$ (for any $j > 2^{4k} e \ln(1/\epsilon)$) is at most $\left(\frac{2^{4k} e \ln(1/\epsilon)}{j}\right)^j$.*

Proof: By Lemma 30, $T_A := \sum_{i=1}^{t} \Pr[T_i] < 2^{4k} \ln(1/\epsilon)$. The probability that some set of j A_i's is satisfied is at most $\sum_{S \subseteq [t], |S|=j} \Pr[\wedge_{i \in S} A_i]$. Applying Claim 27, we have:

$$\sum_{S \subseteq [t], |S|=j} \Pr[\wedge_{i \in S} A_i] \leq \sum_{S \subseteq [t], |S|=j} \prod_{i \in S} \Pr[T_i].$$

The arithmetic-geometric mean inequality shows that this quantity is maximized when all $\Pr[T_i]$ are equal, hence:

$$\sum_{S \subseteq [t], |S|=j} \prod_{i \in S} \Pr[T_i] \leq \binom{t}{j} \left(\frac{T_A}{t}\right)^j \leq \left(\frac{e T_A}{j}\right)^j \leq \left(\frac{2^{4k} e \ln 1/\epsilon}{j}\right)^j$$

■

We can now show that Mansour's conjecture holds for read-k DNF formulas with any constant k.

Theorem 32 *Let $f : \{0,1\}^n \to \{0,1\}$ be any read-k DNF formula with t terms. Then there is a polynomial $q_{f,d}$ with $\|q_{f,d}\|_1 = t^{O(2^{4k} \log 1/\epsilon)}$ and $\mathbf{E}[(f - q_{f,d})^2] \leq \epsilon$ for all $\epsilon > 0$.*

Proof: If $\Pr[f = 1] > 1 - \epsilon$, the constant 1 is a suitable polynomial. Let g be the DNF formula f after dropping terms of length greater than $w := \log(2t/\epsilon)$. (This only changes the probability by $\epsilon/2$.) Let $d := \lceil 4e^3 2^{4k} \ln(2/\epsilon) \rceil$ and $q_{g,d}$ be as defined at the beginning of Section 5. Lemma 26 tells us that $\|q_{g,d}\|_1 \leq t^{O(2^{4k} \log 1/\epsilon)}$, and Lemma 31 combined with Lemma 18 tells us that $\mathbf{E}[(g - q_{g,d})^2] \leq \epsilon/2$. ■

6 Pseudorandomness

De et al. (2009) recently improved long-standing pseudorandom generators against DNF formulas.

Definition 33 *A probability distribution X over $\{0,1\}^n$ ϵ-fools a real function $f : \{0,1\}^n \to \mathbb{R}$ if*

$$|\mathbf{E}[f(X)] - \mathbf{E}[f(U_n)]| \leq \epsilon.$$

If $\mathcal{C}$ is a class of functions, then we say that X ϵ-fools $\mathcal{C}$ if X ϵ-fools every function $f \in \mathcal{C}$.

We say a probability distribution X over $\{0,1\}^n$ is ϵ-biased *if it ϵ-fools the character function χ_S for every $S \subseteq [n]$.*

De et al. (2009) observed that the result of Bazzi (2007) implied a pseudorandom generator that ϵ-fools t-term DNF formulas over n variables with seed length $O(\log n \cdot \log^2(t/\beta))$, which already improves the long-standing upper bound of $O(\log^4(tn/\epsilon))$ of Luby et al. (1993). They go on to show a pseudorandom generator with seed length $O(\log n + \log^2(t/\epsilon) \log\log(t/\epsilon))$.

They prove that a sufficient condition for a function f to be ϵ-fooled by an ϵ-biased distribution is that the function be "sandwiched" between two bounded real-valued functions whose Fourier transform has small ℓ_1 norm:

Lemma 34 (Sandwich Bound (De et al., 2009)) *Suppose $f, f_\ell, f_u : \{0,1\}^n \to \mathbb{R}$ are three functions such that for every $x \in \{0,1\}^n$, $f_\ell(x) \leq f(x) \leq f_u(x)$, $\mathbf{E}[f_u(U_n)] - \mathbf{E}[f(U_n)] \leq \epsilon$, and $\mathbf{E}[f(U_n)] - \mathbf{E}[f_\ell(U_n)] \leq \epsilon$. Let $L = \max(\|f_\ell\|_1^{\neq\emptyset}, \|f_u\|_1^{\neq\emptyset})$. Then any β-biased probability distribution $(\epsilon + \beta L)$-fools f.*

Naor and Naor (1993) prove that an ϵ-biased distribution over n bits can be sampled using a seed of $O(\log(n/\epsilon))$ bits. Using our construction from Section 4, we show that random DNF formulas are ϵ-fooled by a pseudorandom generator with seed length $O(\log n + \log(t)\log(1/\epsilon))$:

Theorem 35 *Let $f = T_1 \vee \cdots \vee T_t$ be a random DNF formula chosen from $\mathcal{D}_n^t$ for $t = n^{\Theta(1)}$. For $1 \leq d \leq t$, with probability $1 - 1/n^{\Omega(\log t)}$ over the choice of f, β-biased distributions $O(2^{-\Omega(d)} + \beta t^d)$-fool f. In particular, we can ϵ-fool most $f \in \mathcal{D}_n^t$ by a $t^{-O(\log(1/\epsilon)}$-biased distribution.*

Proof: Let d^+ be the first odd integer greater than d, and let d^- be the first even integer greater than d. Let $f_u = p_{f,d^+}$ and $f_\ell = p_{f,d^-}$ (where $p_{f,d}$ is defined as in Section 3). By Lemma 16, the ℓ_1-norms of f_u and f_ℓ are $t^{O(d)}$. By Fact 15, we know that $P_{d^+}(y) = \binom{y-1}{d} + 1 > 1$ and $P_{d^-}(y) = -\binom{y-1}{d} + 1 \leq 0$ for $y \in [t] \setminus [d]$, hence:

$$\mathbf{E}[f_u(U_n)] - \mathbf{E}[f(U_n)] = \sum_{j=d+1}^{t} \left(\binom{j-1}{d} + 1 - 1 \right) \Pr[y_f = j],$$

which with probability $1 - 1/n^{\Omega(\log t)}$ over the choice of f is at most $2^{-\Omega(d)}$ by the analysis in Lemma 18. The same analysis applies for f_ℓ, thus applying Lemma 34 gives us the theorem. ■

De et al. (2009) match our bound for random DNF formulas for the special case of read-once DNF formulas. Using our construction from Section 5 and a similar proof as the one above, we can show that monotone read-k formulas are ϵ-fooled by a pseudorandom generator with seed length $O(\log n + \log(t)\log(1/\epsilon))$.

Theorem 36 *Let $f = T_1 \vee \cdots \vee T_t$ be a read-k DNF formula for constant k. For $1 \leq d \leq t$, β-biased distributions $O(2^{-\Omega(d)} + \beta t^d)$-fool f. In particular, we can ϵ-fool read-k DNF formulas by a $t^{-O(\log(1/\epsilon))}$-biased distribution.*

Acknowledgments. Thanks to Sasha Sherstov for important contributions at an early stage of this work. We would also like to thank Omid Etesami for pointing out some errors in a previous version of this work, including a crucial flaw in the proof of the read-k case. We also thank him for pointing out to us that our proof for the monotone read-k case extends to the non-monotone case (through Lemma 28). Lemma 28 is due to Omid and James Cook.

References

Alon, N., & Spencer, J. H. (2000). *The probabilistic method*. Hoboken, NJ: Wiley-Interscience. 2nd edition edition.

Bazzi, L. (2007). Polylogarithmic independence can fool DNF formulas. *Proc. 48th IEEE Symposium on Foundations of Computer Science (FOCS)* (pp. 63–73).

De, A., Etesami, O., Trevisan, L., & Tulsiani, M. (2009). *Improved pseudorandom generators for depth 2 circuits* (Technical Report 141). Electronic Colloquium on Computational Complexity (ECCC).

Etesami, O., & Cook, J. (2010). personal communication.

Gopalan, P., Kalai, A., & Klivans, A. R. (2008a). A query algorithm for agnostically learning DNF? *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008* (pp. 515–516). Omnipress.

Gopalan, P., Kalai, A. T., & Klivans, A. R. (2008b). Agnostically learning decision trees. *Proceedings of the 40th Annual ACM Symposium on Theory of Computing, Victoria, British Columbia, Canada, May 17-20, 2008* (pp. 527–536). ACM.

Graham, R. L., Knuth, D. E., & Patashnik, O. (1994). *Concrete mathematics: A foundation for computer science*. Addison-Wesley.

Hancock, T., & Mansour, Y. (1991). Learning monotone k-μ DNF formulas on product distributions. *Proc. of the 4th Annual Conference on Computational Learning Theory (COLT)* (pp. 179–183).

Håstad, J. (1986). *Computational limitations for small depth circuits*. MIT Press.

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

Jackson, J. C., Lee, H. K., Servedio, R. A., & Wan, A. (2008). Learning random monotone DNF. *12th Intl. Workshop on Randomization and Computation (RANDOM)* (pp. 483–497). Springer-Verlag.

Jackson, J. C., & Servedio, R. A. (2005). On learning random DNF formulas under the uniform distribution. *9th Intl. Workshop on Randomization and Computation (RANDOM)* (pp. 342–353). Springer-Verlag.

Kalai, A., Klivans, A., Mansour, Y., & Servedio, R. (2008). Agnostically learning halfspaces. *SIAM Journal on Computing*, *37*, 1777–1805.

Kleitman, D. J. (1966). Families of non-disjoint subsets. *Journal of Combinatorial Theory*, *1*, 153–155.

Kushilevitz, E., & Mansour, Y. (1993). Learning decision trees using the Fourier spectrum. *SIAM Journal on Computing*, *22*, 1331–1348. Prelim. ver. in *Proc. of STOC'91*.

Linial, N., Mansour, Y., & Nisan, N. (1993). Constant depth circuits, Fourier transform and learnability. *Journal of the ACM*, *40*, 607–620.

Luby, M., Velickovic, B., & Wigderson, A. (1993). Deterministic approximate counting of depth-2 circuits. *ISTCS 1993* (pp. 18–24).

Mansour, Y. (1994). *Learning Boolean functions via the Fourier transform*, 391–424. Kluwer Academic Publishers.

Mansour, Y. (1995). An $O(n^{\log\log n})$ learning algorithm for DNF under the uniform distribution. *Journal of Computer and System Sciences*, *50*, 543–550. Prelim. ver. in *Proc. of COLT'92*.

Naor, J., & Naor, M. (1993). Small-bias probability spaces: Efficient constructions and applications. *SIAM Journal on Computing*, *22*, 838–856.

Razborov, A. (2008). *A simple proof of Bazzi's theorem* (Technical Report 81). Electronic Colloquium on Computational Complexity (ECCC).

Sellie, L. (2008). Learning random monotone DNF under the uniform distribution. *Proc. of the 21th Annual Conference on Computational Learning Theory (COLT)* (pp. 181–192).

Sellie, L. (2009). Exact learning of random DNF over the uniform distribution. *Proc. 41st Annual ACM Symposium on Theory of Computing (STOC)* (pp. 45–54).