On Learning Random DNF Formulas under the Uniform Distribution

Jeffrey C. Jackson*  
Dept. of Math. and Computer Science  
Duquesne University  
Pittsburgh, PA 15282  
jackson@mathcs.duq.edu

Rocco A. Servedio†  
Dept. of Computer Science  
Columbia University  
New York, NY 10027, USA  
rocco@cs.columbia.edu

April 13, 2005

Abstract

We study the average-case learnability of DNF formulas in the model of learning from uniformly distributed random examples. We define a natural model of random monotone DNF formulas and give an efficient algorithm which with high probability can learn, for any fixed constant $\gamma > 0$, a random $t$-term monotone DNF for any $t = O(n^{2-\gamma})$. We also define an analogous model of random nonmonotone DNF and give an efficient algorithm which with high probability can learn a random $t$-term DNF for any $t = O(n^{3/2-\gamma})$. These are the first known algorithms that can successfully learn a broad class of polynomial-size DNF in a reasonable average-case model of learning from random examples.

*This material is based upon work supported by the National Science Foundation under Grant No. CCR-0209064.
†This material is based upon work supported by NSF Early Career Development (CAREER) Grant CCF-0347282.
1 Introduction

1.1 Motivation and Background. A disjunctive normal form formula, or DNF, is an AND of ORs of Boolean literals. A question that has been open since Valiant’s initial paper on computational learning theory [23] is whether or not efficient algorithms exist for learning polynomial size DNF formulas in various learning models. The only positive result to date is the Harmonic Sieve [11], which is a membership-query algorithm that efficiently learns DNF with respect to the uniform distribution (and certain related distributions). The approximating function produced by the Sieve is not itself a DNF; thus, the Sieve is an improper learning algorithm.

There has been little progress on polynomial-time algorithms for learning arbitrary DNF since the discovery of the Sieve. There are two obvious relaxations of the uniform-plus-membership model that can be pursued: learn with respect to arbitrary distributions using membership queries, and learn with respect to uniform without membership queries. However, given standard cryptographic assumptions, the first direction is essentially as difficult as showing that DNF is learnable with respect to arbitrary distributions without membership queries [3]. On the other hand, there are substantial known obstacles to learning DNF in the second model of uniform distribution without membership queries. In particular, Blum et al. [7] showed that no algorithm which can be recast as a Statistical Query algorithm can learn arbitrary polynomial-size DNF under the uniform distribution in $n^{O(\log n)}$ time. Since nearly all non-membership learning algorithms can be recast as SQ algorithms [15], a major conceptual shift seems necessary to obtain an algorithm for efficiently learning arbitrary DNF formulas from uniform examples alone.

An apparently simpler question is whether or not we can efficiently learn monotone DNF formulas. Angluin showed early on that monotone DNF can be properly learned with respect to arbitrary distributions using membership queries [2]. It has also long been known that it is as hard to learn monotone DNF with respect to arbitrary distributions without membership queries as it is to learn arbitrary DNF in the same model [16]. This leaves the following enticing question (posed in, e.g., [14], [6] and [5]): are monotone DNF efficiently learnable from uniform examples alone?

In 1990, Verbeurg [24] gave an algorithm that can properly learn any poly(n)-size (arbitrary) DNF from uniform examples in time $n^{O(\log n)}$. More recently, an algorithm has been developed that learns any $2^{\sqrt{\log n}}$-term monotone DNF in poly(n) time [22]. However, despite significant interest in the problem, no algorithm faster than that of [24] is known for learning arbitrary poly(n)-size monotone DNF from uniform examples, and no known hardness result precludes such an algorithm (the SQ result of Blum et al. is at its heart a hardness result for low-degree parity functions).

Since worst-case versions of several DNF learning problems have remained stubbornly open for a decade or more, it is natural to ask about DNF learning from an average-case perspective, i.e., about learning random DNF formulas. In fact, this question has been considered before: Aizenstein & Pitt [1] were the first to ask whether random DNF formulas are efficiently learnable. They proposed a model of random DNF in which each of the $t$ terms is selected independently at random from all possible terms, and gave a membership and equivalence query algorithm which with high probability learns a random DNF generated in this way. However, as noted in [1], a limitation of this model is that with very high probability all terms will have length $\Omega(n)$. The learning algorithm itself becomes quite simple in this situation. Thus, while this is a “natural” average-case DNF model, from a learning perspective it is not a particularly interesting model. To address this deficiency, they also proposed another natural average-case model which is parameterized by the expected length $k$ of each term as well as the number of independent terms $t$, but left open the question of whether or not random DNF can be efficiently learned in such a model.

1.2 Our results. We consider an average-case DNF model very similar to the latter Aizenstein and Pitt model, although we simplify slightly by assuming that $k$ represents a fixed term length
rather than an expected length. We show that, in the model of learning from uniform random examples only, random monotone DNF are properly and efficiently learnable for many interesting values of $k$ and $t$. In particular, for $t = O(n^{2/3})$ where $\gamma > 0$, and for $k = \log t$, our algorithm can achieve any error rate $\epsilon > 0$ in $\poly(n, 1/e)$ time with high probability (over both the selection of the target DNF and the selection of examples). In addition, we obtain slightly weaker results for arbitrary DNF: our algorithm can properly and efficiently learn random $t$-term DNF for $t$ such that $t = O(n^{2/3})$. This algorithm cannot achieve arbitrarily small error but can achieve error $\epsilon = o(1)$ for any $t = \omega(1)$. For detailed result statements see Theorems 13 and 24.

While our results would clearly be stronger if they held for any $t = \poly(n)$ rather than the specific polynomials given, they are a marked advance over the previous state of affairs for DNF learning. (Recall that in the standard worst-case model, $\poly(n)$-time uniform-distribution learning of $t(n)$-term DNF for any $t(n) = \omega(1)$ is an open problem with an associated cash prize [4].)

We note that taking $k = \log t$ is a natural choice when learning with respect to the uniform distribution. (We actually allow a somewhat more general choice of $k$.) This choice leads to target DNFs that, with respect to uniform, are roughly balanced (0 and 1 values are equally likely). From a learning perspective balanced functions are generally more interesting than unbalanced functions, since a constant function is trivially a good approximator to a highly unbalanced function.

Our results shed some light on which cases are not hard in the worst-case model. While “hard” cases were previously known for arbitrary DNF [4], our findings may be particularly helpful in guiding future research on monotone DNF. In particular, our algorithm learns any monotone DNF which (i) is near-balanced, (ii) has every term uniquely satisfied with reasonably high probability, (iii) has every pair of terms jointly satisfied with much smaller probability, and (iv) has no variable appearing in significantly more than a $1/\sqrt{t}$ fraction of the $t$ terms (this is made precise in Lemma 9). So in order to be “hard,” a monotone DNF must violate one or more of these criteria.

Our algorithms work in two stages: they first identify pairs of variables which co-occur in some term of the target DNF, and then use these pairs to reconstruct terms via a specialized clique-finding algorithm. For monotone DNF we can with high probability determine for every pair of variables whether or not the pair co-occurs in some term. For nonmonotone DNF, with high probability we can identify most pairs of variables which co-occur in some term; as we show, this enables us to learn to fairly (but not arbitrarily) high accuracy.

We give preliminaries in Section 2. Sections 3 and 4 contain our results for monotone and nonmonotone DNF respectively. Section 5 concludes.

2 Preliminaries

We first describe our models of random monotone and nonmonotone DNF. Let $\mathcal{M}_{n,k}^t$ be the probability distribution over monotone $t$-term DNF induced by the following random process: each term is independently and uniformly chosen at random from all $\binom{n}{k}$ monotone ANDs of size exactly $k$ over variables $v_1, \ldots, v_n$. For nonmonotone DNF, we write $\mathcal{D}_{n,k}^t$ to denote the following distribution over $t$-term DNF: first a monotone DNF is selected from $\mathcal{M}_{n,k}$, and then each occurrence of each variable in each term is independently negated with probability $1/2$. (Equivalently, a draw from $\mathcal{D}_{n,k}^t$ is done by independently selecting $t$ terms from the set of all terms of length exactly $k$).

Given a Boolean function $\phi : \{0, 1\}^n \to \{0, 1\}$, we write $\Pr[\phi]$ to denote $\Pr_{x \sim \mathcal{U}_n}[\phi(x) = 1]$, where $\mathcal{U}_n$ denotes the uniform distribution over $\{0, 1\}^n$. We write $\log log$ to denote $\log 2$ and $\log$ to denote natural log. Many of our proofs use tail bounds such as Chernoff bounds and McDiarmid’s inequality: these bounds are given in Appendix A.

In the uniform distribution learning model which we consider, the learner is given a source of labeled examples $(x, f(x))$ where each $x$ is uniformly drawn from $\{0, 1\}^n$ and $f$ is the unknown function to be learned. The goal of the learner is to efficiently construct a hypothesis $h$ which with
high probability (over the choice of labeled examples used for learning) has low error relative to \( f \) under the uniform distribution, i.e. \( \Pr_{x \sim U_n}[h(x) \neq f(x)] \leq \epsilon \) with probability \( 1 - \delta \). This model has been intensively studied in learning theory, see e.g. [10, 9, 12, 18, 19, 21, 22, 24]. In our average case framework, the target function \( f \) will be drawn randomly from either \( \mathcal{M}_{n,k}^t \) or \( \mathcal{P}_{n,k}^t \), and (as in [13]) our goal is to construct a low-error hypothesis \( h \) for \( f \) with high probability over both the random examples used for learning and the random draw of \( f \).

3 Learning Random Monotone DNF

3.1 Interesting Parameter Settings. Consider a random draw of \( f \) from \( \mathcal{M}_{n,k}^t \). It is intuitively clear that if \( t \) is too large relative to \( k \) then a random \( f \in \mathcal{M}_{n,k}^t \) will likely have \( \Pr[f] \approx 1 \); similarly if \( t \) is too small relative to \( k \) then a random \( f \in \mathcal{M}_{n,k}^t \) will likely have \( \Pr[f] \approx 0 \). Such cases are not very interesting from a learning perspective since a trivial algorithm can learn to high accuracy. We are thus led to the following definition:

**Definition 1** For \( 0 < \alpha \leq 1/2 \), a pair of values \((k, t)\) is said to be monotone \( \alpha \)-interesting if \( \alpha \leq E_{f \in \mathcal{M}_{n,k}^t}[\Pr[f]] \leq 1 - \alpha \).

Throughout the paper we will assume that \( 0 < \alpha \leq 1/2 \) is a fixed constant independent of \( n \) and that \( t \leq p(n) \), where \( p(\cdot) \) is a fixed polynomial (and we will also make assumptions about the degree of \( p \)). The following lemma proved in Appendix B gives necessary conditions for \((k, t)\) to be monotone \( \alpha \)-interesting: (As Lemma 2 indicates, we may always think of \( k \) as being roughly \( \log t \))

**Lemma 2** If \((k, t)\) is monotone \( \alpha \)-interesting then \( \alpha 2^k \leq t \leq 2^{k+1} \ln \frac{2}{\alpha} \).

3.2 Properties of Random Monotone DNF. Throughout the rest of Section 3 we assume that \( \alpha > 0 \) is fixed and \((k, t)\) is a monotone \( \alpha \)-interesting pair where \( t = O(n^{2-\gamma}) \) for some \( \gamma > 0 \). In this section we develop some useful lemmas regarding \( \mathcal{M}_{n,k}^t \). All proofs are given in Appendix C.

Let \( f^i \) denote the projected function obtained from \( f \) by first removing term \( T_i \) from the monotone DNF for \( f \) and then restricting all of the variables which were present in term \( T_i \) to 1. For \( \ell \neq i \) we write \( T_{\ell}^i \) to denote the term obtained by setting all variables in \( T_i \) to 1 in \( T_\ell \), i.e. \( T_{\ell}^i \) is the term in \( f^i \) corresponding to \( T_\ell \). Note that if \( T_{\ell}^i \neq T_i \) then \( T_{\ell}^i \) is smaller than \( T_i \).

The following two lemmas show that each variable appears in a limited number of terms and that therefore not too many terms \( T_{\ell}^i \) in \( f^i \) are smaller than their corresponding terms \( T_i \) in \( f \). In these and later lemmas, “\( n \) sufficiently large” means that \( n \) is larger than a constant which depends on \( \alpha \) but not on \( k \) or \( t \).

**Lemma 3** For \( n \) sufficiently large, with probability at least \( 1 - \delta_{\text{many}} := 1 - n \left( \frac{e^{k^{3/2} \log t}}{n^2 - 2^{k^2}} \right)^2 2^{-1} \alpha^2 / \sqrt{\log t} \) over the random draw of \( f \) from \( \mathcal{M}_{n,k}^t \) we have that every variable \( v_j \), \( 1 \leq j \leq n \), appears in at most \( 2^{k-1} \alpha^2 / \sqrt{\log t} \) terms of \( f \).

Note that since \((k, t)\) is a monotone \( \alpha \)-interesting pair and \( t = O(n^{2-\gamma}) \) for some fixed \( \gamma > 0 \), for sufficiently large \( n \) this probability bound is non-trivial.

**Lemma 4** For \( n \) sufficiently large, with probability at least \( 1 - \delta_{\text{small}} := 1 - \left( \frac{k \log t}{n^2} \right)^{2k}/(\log t) \) over the random draw of \( f \) from \( \mathcal{M}_{n,k}^t \) we have that for all \( 1 \leq i \leq t \) at most \( 2^k / \log t \) terms \( T_{\ell}^i \) with \( \ell \neq i \) in the projection \( f^i \) are smaller than the corresponding terms \( T_i \) in \( f \).

There is probably little overlap between any pair of terms in \( f \):
Lemma 5 With probability at least \(1 - \delta_{\text{shared}} := 1 - t^2 \left( \frac{\log t}{n} \right)^2 \) over the random draw of \( f \) from \( \mathcal{M}_n^{t,k} \), for all \( 1 \leq i, j \leq t \) no set of \( \log t \) or more variables belongs to two distinct terms \( T_i \) and \( T_j \) in \( f \).

Using the preceding lemmas, we can show that for \( f \) drawn from \( \mathcal{M}_n^{t,k} \), with high probability each term is “uniquely satisfied” by a noticeable fraction of assignments. More precisely, we have:

Lemma 6 For \( n \) sufficiently large, with probability at least \(1 - \delta_{\text{small}} - \delta_{\text{shared}} := 1 - \delta_{\text{sat}}\) over the random draw of \( f \) from \( \mathcal{M}_n^{t,k} \), \( f \) is such that for all \( i = 1, \ldots, t \) we have \( \Pr_x [T_i \text{ is satisfied by } x] \geq \frac{a^3}{2^{t-2}} \).

On the other hand, we can upper bound the probability that two terms of a random DNF \( f \) will be satisfied simultaneously:

Lemma 7 With probability at least \(1 - \delta_{\text{shared}}\) over the random draw of \( f \) from \( \mathcal{M}_n^{t,k} \), for all \( 1 \leq i < j \leq t \) we have \( \Pr [T_i \land T_j] \leq \frac{\log t}{2^t} \).

3.3 Identifying cooccurring variables. In this section we show how to identify pairs of variables \((v_i, v_j)\) which cooccur in some term of \( f \).

First, some notation. Given a monotone DNF \( f \) over variables \( v_1, \ldots, v_n \), define DNF formulas \( g_{ss}, g_{1s}, g_{s1} \) and \( g_{11} \) over variables \( v_3, \ldots, v_n \) as follows:
- \( g_{ss} \) is the disjunction of the terms in \( f \) that contain neither \( v_1 \) nor \( v_2 \);
- \( g_{1s} \) is the disjunction of the terms in \( f \) that contain \( v_1 \) but not \( v_2 \) (but with \( v_1 \) removed from each of these terms);
- \( g_{s1} \) is defined similarly as the disjunction of the terms in \( f \) that contain \( v_2 \) but not \( v_1 \) (but with \( v_2 \) removed from each of these terms);
- \( g_{11} \) is the disjunction of the terms in \( f \) that contain both \( v_1 \) and \( v_2 \) (with both variables removed from each term).

We thus have \( f = g_{ss} \lor (v_1 g_{1s}) \lor (v_2 g_{s1}) \lor (v_1 v_2 g_{11}) \). Note that any of \( g_{ss}, g_{1s}, g_{s1}, g_{11} \) may be an empty disjunction which is identically false.

We can empirically estimate each of the following using uniform random examples \((x, f(x))\):

\[
p_{00} := \Pr_x [g_{ss}] = \Pr_{x \in U_n} [f(x) = 1 | x_1 = x_2 = 0]
\]
\[
p_{01} := \Pr_x [g_{ss} \lor g_{1s}] = \Pr_{x \in U_n} [f(x) = 1 | x_1 = 0, x_2 = 1]
\]
\[
p_{10} := \Pr_x [g_{ss} \lor g_{s1}] = \Pr_{x \in U_n} [f(x) = 1 | x_1 = 1, x_2 = 0]
\]
\[
p_{11} := \Pr_x [g_{ss} \lor g_{1s} \lor g_{s1} \lor g_{11}] = \Pr_{x \in U_n} [f(x) = 1 | x_1 = 1, x_2 = 1].
\]

It is clear that \( g_{11} \) is nonempty if and only if \( v_1 \) and \( v_2 \) cooccur in some term of \( f \); thus we would ideally like to obtain \( \Pr_{x \in U_n} [g_{11}] \). While we cannot obtain this probability from \( p_{00}, p_{01}, p_{10} \) and \( p_{11} \), the following lemma, proved in Appendix D, shows that we can estimate a related quantity:

Lemma 8 Let \( P \) denote \( p_{11} - p_{10} - p_{01} + p_{00} \). Then \( P = \Pr [g_{11} \land \overline{g}_{1s} \land \overline{g}_{s1} \land \overline{g}_{ss}] - \Pr [g_{1s} \land g_{s1} \land \overline{g}_{ss}] \).

More generally, let \( P_{ij} \) be defined as \( P \) but with \( v_i, x_i, v_j, \) and \( x_j \) substituted for \( v_1, x_1, v_2, \) and \( x_2 \), respectively, throughout the definitions of the \( g \)’s and \( p \)’s above. The following lemma shows that, for most random choices of \( f \), for all \( 1 \leq i, j \leq n \), the value of \( P_{ij} \) is a good indicator of whether or not \( v_i \) and \( v_j \) cooccur in some term of \( f \).
Lemma 9 For $n$ sufficiently large and $t \geq 4$, with probability at least $1 - \delta_{\text{usat}} - \delta_{\text{shared}} - \delta_{\text{many}}$ over the random draw of $f$ from $\mathcal{M}_n^{k}$, we have that for all $1 \leq i, j \leq n$ (i) if $v_i$ and $v_j$ do not cooccur in some term of $f$ then $P_{ij} \leq 0$; (ii) if $v_i$ and $v_j$ do cooccur in some term of $f$ then $P_{ij} \geq \frac{a^2}{t^2}$.

Proof: Part (i) holds for any monotone DNF by Lemma 8. For (ii), we first note that with probability at least $1 - \delta_{\text{many}} - \delta_{\text{small}} - \delta_{\text{usat}}$, a randomly chosen $f$ has all the following properties:

1. Each term in $f$ is uniquely satisfied with probability at least $\alpha^3/2^{k+2}$ (by Lemma 6);

2. Each pair of terms $T_i$ and $T_j$ in $f$ are both satisfied with probability at most $\log t/2^{2k}$ (by Lemma 7); and

3. Each variable in $f$ appears in at most $2^{k-1}\alpha^2/\sqrt{\log t}$ terms (by Lemma 3).

We call such an $f$ well-behaved. For the sequel, assume that $f$ is well-behaved and also assume without loss of generality that $i = 1$ and $j = 2$. We consider separately the two probabilities $\rho_1 = \Pr[g_{11} \land \overline{g}_{a1} \land \overline{g}_{b1} \land \overline{g}_{a_b}]$ and $\rho_2 = \Pr[g_{a1} \land g_{a_b} \land \overline{g}_{b1}]$ whose difference defines $P_{12} = P$. By property (1) above, $\rho_1 \geq \alpha^3/2^{k+2}$, since each instance $x$ that uniquely satisfies a term $T_j$ in $f$ containing both $v_1$ and $v_2$ also satisfies $g_{11}$ while falsifying all of $g_{a1}$, $g_{a_b}$, and $g_{b1}$. Since $(k, t)$ is monotone $\alpha$-interesting, this implies that $P_{12} \geq \alpha^4/4t$. On the other hand, clearly $\rho_2 \leq \Pr[g_{a1} \land g_{a_b}]$.

By property (2) above, for any pair of terms consisting of one term from $g_{a1}$ and the other from $g_{a_b}$, the probability that both terms are satisfied is at most $\log t/2^{2k}$. Since each of $g_{a1}$ and $g_{a_b}$ contains at most $2^{k-1}\alpha^2/\sqrt{\log t}$ terms by property (3), by a union bound we have $\rho_2 \leq \alpha^4/(4t \log t)$, and the lemma follows given the assumption that $t \geq 4$.

Thus, our algorithm for finding all of the cooccurring pairs of a randomly chosen monotone DNF consists of estimating $P_{ij}$ for each of the $n(n - 1)/2$ pairs $(i, j)$ so that all of our estimates are—with probability at least $1 - \delta$—within an additive factor of $\alpha^4/16t$ of their true values. The reader familiar with Boolean Fourier analysis will readily recognize that $P_{12}$ is just $\hat{f}(110_{n-2})$ and that in general all of the $P_{ij}$ are simply second-order Fourier coefficients. Therefore, by the standard Hoeffding bound, a uniform random sample of size $O(t^2 \ln(n^2/\delta)/\alpha^8)$ is sufficient to estimate all of the $P_{ij}$’s to the specified tolerance with overall probability at least $1 - \delta$. We thus have:

Theorem 10 For $n$ sufficiently large and any $\delta > 0$, with probability at least $1 - \delta_{\text{usat}} - \delta_{\text{shared}} - \delta_{\text{many}} - \delta$ over the choice of $f$ from $\mathcal{M}_n^{k}$ and the choice of random examples, our algorithm runs in $O(n^2 \log(n/\delta))$ time and identifies exactly those pairs $(v_i, v_j)$ which cooccur in some term of $f$.

3.4 Forming a hypothesis from pairs of cooccurring variables. Here we show how to construct an accurate DNF hypothesis for a random $f$ drawn from $\mathcal{M}_n^{k}$.

Identifying all $k$-cliques. By Theorem 10, with high probability we have complete information about which pairs of variables $(v_i, v_j)$ cooccur in some term of $f$. We thus may consider the graph $G$ with vertices $v_1, \ldots, v_n$ and edges for precisely those pairs of variables $(v_i, v_j)$ which cooccur in some term of $f$. This graph is a union of $t$ randomly chosen $k$-cliques from $\{v_1, \ldots, v_n\}$ which correspond to the $t$ terms in $f$. We will show how to efficiently identify (with high probability over the choice of $f$ and random examples of $f$) all of the $k$-cliques in $G$. Once these $k$-cliques have been identified, as we show later it is easy to construct an accurate DNF hypothesis for $f$.

The following lemma (whose proof is in Appendix E) shows that with high probability over the choice of $f$, each pair $(v_i, v_j)$ cooccurs in at most a constant number of terms:

Lemma 11 Fix $1 \leq i < j \leq n$. For any $C \geq 0$ and all sufficiently large $n$, we have $\Pr_{f \in \mathcal{M}_n^{k}}[\text{some pair of variables } (v_i, v_j) \text{ cooccur in more than } C \text{ terms of } f] \leq \delta_C := (\frac{t^2}{n^2})^C$. 

5
By Lemma 11 we know that, for any given pair \((v_i, v_j)\) of variables, with probability at least \(1 - \delta_C\) there are at most \(Ck\) other variables \(v_l\) such that \((v_i, v_j, v_l)\) all cooccur in some term of \(f\). Suppose that we can efficiently (with high probability) identify the set \(S_{ij}\) of all such variables \(v_l\). Then we can perform an exhaustive search over all \((k - 2)\)-element subsets \(S'\) of \(S_{ij}\) in at most \((Ck)^k \leq (eC)^k = n^{O(\log C)}\) time, and can identify exactly those sets \(S'\) such that \(S' \cup \{v_i, v_j\}\) is a clique of size \(k\) in \(G\). Repeating this over all pairs of variables \((v_i, v_j)\), we can with high probability identify all \(k\)-cliques in \(G\).

Thus, to identify all \(k\)-cliques in \(G\) it remains only to show that for every pair of variables \(v_i\) and \(v_j\), we can determine the set \(S_{ij}\) of those variables \(v_l\) that cooccur in at least one term with both \(v_i\) and \(v_j\). Assume that \(f\) is such that all pairs of variables cooccur in at most \(C\) terms, and let \(T\) be a set of variables of cardinality at most \(C\) having the following properties:

- In the projection \(f_{T+0}\) of \(f\) in which all of the variables of \(T\) are fixed to 0, \(v_i\) and \(v_j\) do not cooccur in any term; and
- For every set \(T' \subset T\) such that \(|T'| = |T| - 1\), \(v_i\) and \(v_j\) do cooccur in \(f_{T'+0}\).

Then \(T\) is clearly a subset of \(S_{ij}\). Furthermore, if we can identify all such sets \(T\), then their union will be \(S_{ij}\). There are only \(O(n^C)\) possible sets to consider, so our problem now reduces to the following: given a set \(T\) of at most \(C\) variables, determine whether \(v_i\) and \(v_j\) cooccur in \(f_{T+0}\).

The proof of Lemma 9 shows that \(f\) is well-behaved with probability at least \(1 - \delta_{usat} - \delta_{shared} - \delta_{many}\) over the choice of \(f\). Furthermore, if \(f\) is well-behaved then it is easy to see that for every \(|T| \leq C\), \(f_{T+0}\) is also well-behaved, since \(f_{T+0}\) is just \(f\) with \(O(\sqrt{T})\) terms removed (by Lemma 3). That is, removing terms from \(f\) can only make it more likely that the remaining terms are uniquely satisfied, does not change the bound on the probability of a pair of remaining terms being satisfied, and can only decrease the bound on the number of remaining terms in which a remaining variable can appear. Furthermore, Lemma 8 holds for any monotone DNF \(f\). Therefore, if \(f\) is well-behaved then the proof of Lemma 9 also shows that for every \(|T| \leq C\), the \(P_{ij}\)'s of \(f_{T+0}\) can be used to identify the cooccurring pairs of variables within \(f_{T+0}\). It remains to show that we can efficiently simulate a uniform example oracle for \(f_{T+0}\) so that these \(P_{ij}\)'s can be accurately estimated.

In fact, for a given set \(T\), we can simulate a uniform example oracle for \(f_{T+0}\) by filtering the examples from the uniform oracle for \(f\) so that only examples setting the variables in \(T\) to 0 are accepted. Since \(|T| \leq C\), the filter accepts with constant probability at least \(1/2^C\). A Chernoff argument shows that if all \(P_{ij}\)'s are estimated using a single sample of size \(2^{C+10\log \log (2C+2)n^C/\delta})\) (filtered appropriately when needed) then all of the estimates will have the desired accuracy with probability at least \(1 - \delta\). This gives us the following:

**Theorem 12** For \(n\) sufficiently large, any \(\delta > 0\), and any fixed \(C \geq 2\), with probability at least \(1 - \delta_{small} - \delta_{shared} - \delta_{usat} - \delta_C - \delta\) over the random draw of \(f\) from \(M_{n,k}\) and the choice of random examples, all of the \(k\)-cliques of the graph \(G\) can be identified in time \(n^{O(C)}k^2\log(n/\delta)\).

**The main learning result for monotone DNF.** We now have a list \(T'_1, \ldots, T'_N\) (with \(N = O(n^C)\)) of length-\(k\) monotone terms which contains all \(t\) true terms \(T_1, \ldots, T_t\) of \(f\). Now note that the target function \(f\) is simply an OR of some subset of these \(N\) “variables” \(T_1, \ldots, T_N\), so the standard elimination algorithm for learning disjunctions (see e.g. Chapter 1 of [17]) can be used to PAC learn the target function.

Call the above described entire learning algorithm \(A\). In summary, we have proved the following:

**Theorem 13** Fix \(\gamma, \alpha > 0\) and \(C \geq 2\). Let \((k, t)\) be a monotone \(\alpha\)-interesting pair. For any \(\epsilon > 0\), \(\delta > 0\), and \(t = O(n^{2-\gamma})\), algorithm \(A\) will with probability at least \(1 - \delta_{many} - \delta_{small} - \delta_{usat} - \delta_C - \delta\)
(over a random choice of DNF from $\mathcal{M}_{n,k}^k$ and the randomness of the example oracle) produce a hypothesis $h$ that $\epsilon$-approximates the target with respect to the uniform distribution. Algorithm $A$ runs in time polynomial in $n$, $\log(1/\delta)$, and $1/\epsilon$.

4 Nonmonotone DNF

4.1 Interesting Parameter Settings. As with $\mathcal{M}_{n,k}^k$, we are interested in pairs $(k,t)$ for which $\mathbb{E}_{f \in D_{n,k}^k}[\Pr[f]]$ is between $\alpha$ and $1 - \alpha$:

**Definition 14** For $\alpha > 0$, the pair $(k,t)$ is said to be $\alpha$-interesting if $\alpha \leq \mathbb{E}_{f \in D_{n,k}^k}[\Pr[f]] \leq 1 - \alpha$.

For any fixed $x \in \{0,1\}^n$ we have $\Pr_{f \in D_{n,k}^k}[f(x) = 0] = (1 - \frac{1}{2^k})^t$, and thus by linearity of expectation we have $\mathbb{E}_{f \in D_{n,k}^k}[\Pr[f]] = 1 - (1 - \frac{1}{2^k})^t$; this formula will be useful later.

Throughout the rest of Section 4 we assume that $\alpha > 0$ is fixed and $(k,t)$ is an $\alpha$-interesting pair where $t = O(n^{3/2}\gamma)$ for some $\gamma > 0$.

4.2 Properties of Random DNF. In this section we develop analogues of Lemmas 6 and 7 for $D_{n,k}^k$. The $D_{n,k}^k$ analogue of Lemma 7 follows directly from the proof of Lemma 7, and we have:

**Lemma 15** With probability at least $1 - \delta_{\text{shared}}$, over the random draw of $f$ from $D_{n,k}^k$, for all $1 \leq i < j \leq n$, $\Pr[T_i \cap T_j] \leq \frac{\log t}{2^{kt}}$.

In Appendix F we use McDiarmid's bound to prove a $D_{n,k}^k$ version of Lemma 6:

**Lemma 16** With probability at least $1 - \delta_{\text{unsat}} := 1 - t \left( (t - 1) \left( \frac{k^2}{n} \log t \right) \right)$, a random $f$ drawn from $D_{n,k}^k$ is such that for each $i = 1, \ldots, t$, we have $P_i = \Pr_x[T_i \text{ is satisfied by } x]$ but no other $T_j$ is satisfied by $x$.

4.3 Identifying (most pairs of) cooccurring variables. Recall that in Section 3.3 we partitioned the terms of our monotone DNF into four disjoint groups depending on what subset of $\{v_1, v_2\}$ was present in each term. Now, in the nonmonotone case, we will partition the terms of $f$ into nine disjoint groups depending on whether each of $v_1, v_2$ is unnegated, negated, or absent:

$$f = g_{ss} \lor (v_1g_{1s}) \lor (v_1g_{1a}) \lor (v_2g_{2s}) \lor (v_1v_2g_{11}) \lor (v_1v_2g_{10}) \lor (v_1v_2g_{01}) \lor (v_1v_2g_{00})$$

Thus $g_{ss}$ contains those terms of $f$ which contain neither $v_1$ nor $v_2$ in any form; $g_{sa}$ contains the terms of $f$ which contain $\overline{v_1}$ but not $v_2$ in any form (with $\overline{v_1}$ removed from each term); $g_{s1}$ contains the terms of $f$ which contain $v_2$ but not $v_1$ in any form (with $v_2$ removed from each term); and so on. Each $g_{ij}$ is thus a DNF (possibly empty) over literals formed from $v_3, \ldots, v_n$.

We can empirically estimate each of

$$p_{00} := \Pr_{x \in U_n}[g_{ss} \lor g_{sa} \lor g_{s0} \lor g_{s0} = \Pr_{x \in U_n}[f(x) = 1 \mid x_1 = 0, x_2 = 0]$$

$$p_{01} := \Pr_{x \in U_n}[g_{ss} \lor g_{ss} \lor g_{s1} \lor g_{s1} = \Pr_{x \in U_n}[f(x) = 1 \mid x_1 = 0, x_2 = 1]$$

$$p_{10} := \Pr_{x \in U_n}[g_{ss} \lor g_{ss} \lor g_{s0} \lor g_{s1} = \Pr_{x \in U_n}[f(x) = 1 \mid x_1 = 1, x_2 = 0]$$

$$p_{11} := \Pr_{x \in U_n}[g_{ss} \lor g_{ss} \lor g_{s1} \lor g_{s1} = \Pr_{x \in U_n}[f(x) = 1 \mid x_1 = 1, x_2 = 1]$$

It is easy to see that $\Pr[\Delta_{11}]$ is either 0 or else at least $\frac{1}{2^t}$ depending on whether $g_{11}$ is empty or not. Ideally we would like to be able to accurately estimate each of $\Pr[g_{00}], \Pr[g_{01}], \Pr[g_{10}]$
and \( \Pr[g_{11}] \); if we could do this then we would have complete information about which pairs of literals involving variables \( v_1 \) and \( v_2 \) cooccur in terms of \( f \). Unfortunately, the probabilities \( \Pr[g_{00}], \Pr[g_{01}], \Pr[g_{10}] \) and \( \Pr[g_{11}] \) cannot in general be obtained from \( p_{00}, p_{01}, p_{10} \) and \( p_{11} \). However, we will show that we can efficiently obtain some partial information which enables us to learn to fairly high accuracy.

As before, our approach is to accurately estimate the quantity \( P = p_{11} - p_{10} - p_{01} + p_{00} \). We have the following lemmas proved in Appendix G:

**Lemma 17** If all four of \( g_{00}, g_{01}, g_{10} \), and \( g_{11} \) are empty, then \( P \) equals

\[
Pr[g_{1s} \land g_{s0} \land \{ \text{no other } g_{,\} }] + Pr[g_{0s} \land g_{s1} \land \{ \text{no other } g_{,\} }] - Pr[g_{1s} \land g_{s1} \land \{ \text{no other } g_{,\} }] - Pr[g_{0s} \land g_{s0} \land \{ \text{no other } g_{,\} }].
\]

(1)

**Lemma 18** If exactly one of \( g_{00}, g_{01}, g_{10} \), and \( g_{11} \) is nonempty (say \( g_{11} \)), then \( P \) equals (1) plus

\[
Pr[g_{11} \land g_{1s} \land g_{s0} \land \{ \text{no other } g_{,\} }] + Pr[g_{11} \land g_{0s} \land g_{s1} \land \{ \text{no other } g_{,\} }] - Pr[g_{11} \land g_{1s} \land g_{s1} \land \{ \text{no other } g_{,\} }] - Pr[g_{11} \land g_{0s} \land g_{s0} \land \{ \text{no other } g_{,\} }] + Pr[g_{11} \land \{ \text{no other } g_{,\} }] + Pr[g_{11} \land g_{s0} \land \{ \text{no other } g_{,\} }] + Pr[g_{11} \land \{ \text{no other } g_{,\} }].
\]

Using the above two lemmas we can show that the value of \( P \) is a good indicator for distinguishing between all four of \( g_{00}, g_{01}, g_{10}, g_{11} \) being empty versus exactly one of them being nonempty:

**Lemma 19** For \( n \) sufficiently large and \( t \geq 4 \), with probability at least \( 1 - \delta_{\text{usat}} - \delta_{\text{shared}} - \delta_{\text{many}} \) over a random draw of \( f \) from \( \mathcal{D}_{n}^{k,t} \), we have that: (i) if \( v_1 \) and \( v_2 \) do not cooccur in any term of \( f \) then \( P \leq \frac{\alpha^2}{8t} \); (ii) if \( v_1 \) and \( v_2 \) do cooccur in some term of \( f \) and exactly one of \( g_{00}, g_{01}, g_{10} \) and \( g_{11} \) is nonempty, then \( P \geq \frac{3\alpha^2}{16t} \).

**Proof:** With probability at least \( 1 - \delta_{\text{usat}} - \delta_{\text{shared}} - \delta_{\text{many}} \) a randomly chosen \( f \) from \( \mathcal{D}_{n}^{k,t} \) will have all of the following properties:

1. Each term in \( f \) is uniquely satisfied with probability at least \( \alpha/2^{k+1} \) (by Lemma 16);
2. Each variable in \( f \) appears in at most \( 2^{k-1} \alpha^2/\sqrt{t} \log t \) terms (by Lemma 3); and
3. Each pair of terms \( T_i \) and \( T_j \) in \( f \) are both satisfied with probability at most \( \log t/2^{2k} \) (by Lemma 15).

For the sequel assume that we have such an \( f \). We first prove (i) by showing that \( P \)—as represented by (1) of Lemma 17—is at most \( \frac{\alpha^4}{t \log t} \). By property 3 above, for any pair of terms consisting of one term from \( g_{1s} \) and the other from \( g_{s0} \), the probability that both terms are satisfied is at most \( \log t/2^{2k} \). Since each of \( g_{1s} \) and \( g_{s0} \) contains at most \( 2^{k-1} \alpha^2/\sqrt{t} \log t \) terms by property 2, a union bound gives \( \Pr[g_{1s} \land g_{s0} \land \{ \text{no other } g_{,\} }] \leq \Pr[g_{1s} \land g_{s0}] \leq \frac{\alpha^4}{4t \log t} \). A similar argument holds for the three other summands in (1), so \( P \) is at most \( \frac{\alpha^4}{t \log t} \leq \frac{\alpha^2}{8t} \) since \( \alpha \leq 1/2 \) and \( t \geq 4 \).

We now prove (ii). By an argument similar to the above we have that the first six summands in the expression of Lemma 18 are each at most \( \frac{\alpha^4}{4t \log t} \) in magnitude. Now observe that each instance \( x \) that uniquely satisfies a term \( T_j \) in \( f \) containing both \( v_1 \) unnegated and \( v_2 \) unnegated must satisfy \( g_{11} \) and no other \( g_{,\} \). Thus under the conditions of (ii) the last summand in Lemma 18 is at least \( \frac{\alpha^2}{2^{2k+1}} \) by property 1 above, so we have that (ii) is at least \( \frac{\alpha^2}{2^{2k+1}} - \frac{5\alpha^4}{2t \log t} \). Since \( (k,t) \) is \( \alpha \)-interesting we have \( \frac{t}{2^{2k}} \geq \alpha \), and from this and the constant bounds on \( \alpha \) and \( t \) it is easily shown that \( \frac{\alpha^2}{2^{2k+1}} \geq \frac{\alpha^2}{2t} \) and \( \frac{5\alpha^4}{2t \log t} \leq \frac{5\alpha^2}{16t} \), from which the lemma follows.

\[\square\]
It is clear that an analogue of Lemma 19 holds for any pair of variables \( v_i, v_j \) in place of \( v_1, v_2 \). Thus, for each pair of variables \( v_i, v_j \), if we decide whether \( v_i \) and \( v_j \) cooccur (negated or otherwise) in any term on the basis of whether \( P_{ij} \) is large or small, we will err only if two or more of \( g_{00}, g_{01}, g_{10}, g_{11} \) are nonempty.

We now show that for \( f \in D_{n}^{h,k} \), with very high probability there are not too many pairs of variables \( (v_i, v_j) \) which cooccur (with any sign pattern) in at least two terms of \( f \). Note that this immediately bounds the number of pairs \( (v_i, v_j) \) which have two or more of the corresponding \( g_{00}, g_{01}, g_{10}, g_{11} \) nonempty. The following lemma is proved in Appendix H:

**Lemma 20** Let \( d > 0 \) and \( f \in D_{n}^{h,k} \). The probability that more than \((d + 1) t^{2k^4/n^2}\) pairs of variables \( (v_i, v_j) \) each cooccur in two or more terms of \( f \) is at most \( \exp(-d^2 t^{3k^4/n^4}) \).

Taking \( d = n^2 / (t^{5/4} k^4) \) in the above lemma (note that \( d \) is large since \( t^{5/4} = O(n^{15/8}) \)), we have \((d + 1) t^{2k^4/n^2} \leq 2t^{3/4}\) and the failure probability is at most \( \delta_{\text{cooccur}} := \exp(-\sqrt{t}/k^4) \). The results of this section (together with a standard analysis of error in estimating each \( P_{ij} \)) thus yield:

**Theorem 21** For \( n \) sufficiently large and for any \( \delta > 0 \), with probability at least \( 1 - \delta_{\text{cooccur}} - \delta_{\text{sat}} - \delta_{\text{shared}} - \delta_{\text{many}} - \delta \) over the random draw of \( f \) from \( D_{n}^{h,k} \) and the choice of random examples, the above algorithm runs in \( \Theta(n^2 t^{2}\log(n/\delta)) \) time and outputs a list of pairs of variables \( (v_i, v_j) \) such that: (i) if \( (v_i, v_j) \) is in the list then \( v_i \) and \( v_j \) cooccur in some term of \( f \); and (ii) at most \( N_0 = 2t^{3/4} \) pairs of variables \( (v_i, v_j) \) which do cooccur in \( f \) are not on the list.

### 4.4 Reconstructing an accurate DNF hypothesis.

Now we show how to construct a good hypothesis for the target DNF from a list of pairwise cooccurrence relationships as provided by Theorem 21. As in the monotone case, we consider the graph \( G \) with vertices \( v_1, \ldots, v_n \) and edges for precisely those pairs of variables \( (v_i, v_j) \) which cooccur (with any sign pattern) in some term of \( f \). As before this graph is a union of \( t \) randomly chosen \( k \)-cliques \( S_1, \ldots, S_t \) which correspond to the \( t \) terms in \( f \), and as before we would like to find all \( k \)-cliques in \( G \). However, there are two differences now: the first is that instead of having the true graph \( G \), we instead have access only to a graph \( G' \) which is formed from \( G \) by deleting some set of at most \( N_0 = 2t^{3/4} \) edges. The second difference is that the final hypothesis must take the signs of literals in each term into account. To handle these two differences, we use a different reconstruction procedure than we used for monotone DNF in Section 3.4; this reconstruction procedure only works for \( t = O(n^{3/2-\gamma}) \) where \( \gamma > 0 \).

We first show how to identify (with high probability over the choice of \( f \)) the set of all \( k \)-cliques in \( G' \). We then show how to form a DNF hypothesis from the set of all \( k \)-cliques in \( G' \).

We now describe an algorithm which, for \( t = O(n^{3/2-\gamma}) \) with \( \gamma > 0 \), with high probability runs in polynomial time and identifies all the \( k \)-cliques in \( G' \) which contain edge \((v_1, v_2)\). Running the algorithm at most \( tk^2 \) times on all edges in \( G' \) will give us with high probability all the \( k \)-cliques in \( G' \). The algorithm is:

- Let \( \Delta \) be the set of vertices \( v_j \) such that \( v_1, v_2, v_j \) form a triangle in \( G' \). Run a brute-force algorithm to find all \((k - 2)\)-cliques in the graph induced by \( \Delta \).

It is clear that the algorithm finds every \( k \)-clique which contains edge \((v_1, v_2)\). To bound the algorithm’s running time, it suffices to give a high probability bound on the size of \( \Delta \) in the graph \( G \) (clearly \( \Delta \) only shrinks in passing from \( G \) to \( G' \)). The following lemma (proved in Appendix I) gives such a bound:

**Lemma 22** Let \( G \) be a random graph as described above and let \( 0 < \gamma < \frac{1}{t} \). For any \( t = O(n^{3/2-\gamma}) \) and any \( C > 0 \) we have that with probability \( 1 - O\left(\frac{\log \log n}{n^{1/3}}\right) \) the size of \( \Delta \) in \( G \) is at most \( C k \).
By Lemma 22, doing a brute-force search which finds all $k$-cliques in the graph induced by $\Delta$ takes at most $\binom{C_k}{k} \leq (\frac{C_k}{k})^k = (C)^{O(\log n)} = n^{O(\log C)}$ time steps. Thus we can efficiently with high probability identify all the $k$-cliques in $G'$. How many of the “true” cliques $S_1, \ldots, S_t$ in $G$ are not present as $k$-cliques in $G'$? By Lemma 11, with probability at least $1 - t^2 \left( \frac{2}{n} \right)^C$ each edge $(v_i, v_j)$ participates in at most $C$ cliques from $S_1, \ldots, S_t$. Since $G'$ is missing at most $N_0$ edges from $G'$, with probability at least $1 - t^2 \left( \frac{2}{n} \right)^C$ the set of all $k$-cliques in $G'$ is missing at most $CN_0$ “true” cliques from $S_1, \ldots, S_t$.

Summarizing the results of this section so far, we have:

**Theorem 23** Fix $C \geq 2$. Given a DNF formula $f$ drawn from $D_n^t,k$ and a list of pairs of cooccurring variables as described in Theorem 21, with probability at least $1 - 1/n^{\Omega(C)}$ the above procedure runs in $n^{O(\log C)}$ time and constructs a a list $Z_1, \ldots, Z_{N'}$ (where $N' = n^{O(\log C)}$) of $k$-cliques which contains all but at most $CN_0$ of the cliques $S_1, \ldots, S_t$.

We construct a hypothesis DNF from the list $Z_1, \ldots, Z_{N'}$ of candidate $k$-cliques as follows: for each $Z_i$ we form all $2^k$ possible terms which could have given rise to $Z_i$ (corresponding to all $2^k$ sign patterns on the $k$ variables in $Z_i$). We then test each of these $2^kN'$ potential terms against a sample of $M$ randomly drawn negative examples and discard any terms which output 1 on any negative example; the final hypothesis $h$ is the OR of all surviving terms. Any candidate term $T'$ which has $Pr_{x \in U_k}[T'(x) = 1 \& f(x) = 0] \geq \frac{c}{2^{k+1}N'}$ will survive this test with probability at least $\exp(-cM/2^{k+1}N')$. Taking $\epsilon = 1/2^k$ and $M = (1/\epsilon)^{2^k+1}N' \log^2 n$ we have that with probability $1 - 1/n^{\Omega(1)}$ each term in the final hypothesis contributes at most $\epsilon/2^{k+1}N'$ toward the false positive rate of $h$, so with high probability the false positive rate of $h$ is at most $1/2^k$.

The false negative rate of $h$ is at most $\frac{1}{2^k}$ times the number of terms in $f$ which are missing in $h$. Since the above algorithm clearly will not discard any term in $f$ (since such a term will never cause a false negative mistake), we need only bound the number of terms in $f$ which are not among our $2^kN'$ candidates. With probability at least $1 - \delta_{\text{clique}} := 1 - t/\binom{n}{k}$, each true clique $S_1, \ldots, S_t$ in $G$ gives rise to exactly one term of $f$ (the only way this does not happen is if two terms consist of literals over the exact same set of $k$ variables), so Theorem 23 implies that $h$ is missing at most $CN_0$ terms of $f$. Thus the false negative rate is at most $CN_0/2^k \leq 2Ct^{3/4}/2^k = 1/\Omega(t^{1/4})$.

All in all, our main learning result for nonmonotone DNF:

**Theorem 24** Fix $\gamma, \alpha > 0$ and $C \geq 2$. Let $(k, t)$ be a monotone $\alpha$-interesting pair. For $f$ randomly chosen from $D_n^t,k$, with probability at least $1 - \delta_{\text{cooccur}} - \delta_{\text{unsat}} - \delta_{\text{shared}} - \delta_{\text{many}} - \delta_{\text{clique}} = 1/n^{\Omega(C)}$ the above algorithm runs in $\tilde{O}(n^2t^2 + n^{O(\log C)})$ time and outputs a hypothesis $h$ whose error rate relative to $f$ under the uniform distribution is at most $1/\Omega(t^{1/4})$.

One can verify from the definitions of the various $\delta$’s that for any $t = \omega(1)$ as a function of $n$, the failure probability of the algorithm is $o(1)$ and the algorithm learns to accuracy $1 - o(1)$.

5 Discussion and Conclusions

We have shown that several natural models of random DNF formulas can be efficiently learned to high accuracy under the uniform distribution.

Several directions for future work present themselves. We can currently only learn random DNFs with $o(n^{3/2})$ terms ($o(n^2)$ terms for monotone DNF); can stronger results be obtained which hold for all polynomial-size DNF? A natural approach here for learning $n$-term DNF might be to first try to identify all $c$-tuples of variables which cooccur in a term, where $c$ is some constant larger than $c$. Also, our current results for $t = \omega(1)$-term DNF let us learn to some $1 - o(1)$ accuracy but
we cannot yet achieve an arbitrary inverse polynomial error rate for nonmonotone DNF. Finally, another interesting direction is to explore other natural models of random DNF formulas, perhaps by allowing some variation among term sizes or dependencies between terms.

Acknowledgement. Avrim Blum suggested to one of us (JCJ) the basic strategy that learning monotone DNF with respect to uniform might be reducible to finding the cooccurring pairs of variables in the target function.

References


A Tail Bounds

We use the following Chernoff bound [Theorem A.12, Alon & Spencer]: Let $B(t,p)$ denote the binomial distribution with parameter $p$, i.e. a draw from $B(t,p)$ is a sum of $t$ independent $p$-biased 0/1 Bernoulli trials. Then for $\beta > 1$,

$$\Pr_{S \sim B(t,p)} [S \geq \beta pt] \leq \left( e^{\beta - 1} (e/\beta)^{\beta pt}\right)^{pt}.$$

The following bound will also be useful:

**McDiarmid bound** [20]: Let $X_1, \ldots, X_m$ be independent random variables taking values in a set $\Omega$. Let $F: \Omega^m \to \mathbb{R}$ be such that for all $i \in [m]$ we have

$$|F(x_1, \ldots, x_m) - F(x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_m)| \leq c_i$$

for all $x_1, \ldots, x_m$ and $x'_i$ in $\Omega$. Let $\mu = \mathbb{E}[F(X_1, \ldots, X_m)]$. Then for all $\tau > 0$,

$$\Pr [|F(X_1, \ldots, X_m) - \mu| > \tau] < \exp\left(-\tau^2/(c_1^2 + \cdots + c_m^2)\right).$$
B Proof of Lemma 2

Proof of Lemma 2: One side is easy: if \( t < \alpha 2^k \) then each of the \( t \) terms of \( f \) is satisfied by a uniform random example with probability at most \( \alpha/t \), and consequently \( \Pr[f(x) = 1] \leq \alpha \). Note that by our assumptions on \( t \) and \( \alpha \) we thus have that \( k = O(\log n) \) for any monotone \( \alpha \)-interesting pair \( (k, t) \).

We now show that if \( t > 2^{k+1} \log_2 \frac{2}{\alpha} \), then \( \mathbb{E}_{f \in \mathcal{M}_n^{t,k}}[\Pr[f]] > 1 - \alpha \). Let us write \( |x| \) to denote \( x_1 + \cdots + x_n \) for \( x \in \{0,1\}^n \). It is easy to see that \( \Pr[f(x) = 1] \), viewed as a random variable over the choice of \( f \in \mathcal{M}_n^{t,k} \), depends only on the value of \( |x| \). We have

\[
\mathbb{E}_{f \in \mathcal{M}_n^{t,k}}[\Pr[f]] = \sum_{r=0}^{n} \mathbb{E}_{f \in \mathcal{M}_n^{t,k}}[\Pr[f(x) = 1 \mid |x| = r] \cdot \Pr[|x| = r]].
\]

A standard tail bound on the binomial distribution implies that

\[
\Pr_{x \in U_n}[|x| \leq n/2 - \sqrt{n \log(2/\alpha)}] < \alpha/2.
\]

Thus it suffices to show that for any \( x \) with \( |x| \geq n/2 - \sqrt{n \log(2/\alpha)} \), we have \( \Pr_{f \in \mathcal{M}_n^{t,k}}[f(x) = 1] \geq 1 - \alpha/2 \).

Fix an \( x \in \{0,1\}^n \) with \( |x| = w \geq n/2 - \sqrt{n \log(2/\alpha)} \). Let \( T_1 \) be a random monotone term of length \( k \). We have

\[
\Pr[T_1(x) = 1] = \frac{w(w-1)\cdots(w-k+1)}{n(n-1)\cdots(n-k+1)} \geq \frac{1}{2^{k+1}}
\]

where the inequality is implied by our conditions on \( k \) and \( \alpha \). Since the terms of \( f \) are chosen independently, this implies that

\[
\Pr_{f}[f(x) = 0] \leq \left( 1 - \frac{1}{2^{k+1}} \right)^t \leq \exp \left( -\frac{t}{2^{k+1}} \right).
\]

If \( t/2^{k+1} > \ln \frac{2}{\alpha} \) then this bound is at most \( \alpha/2 \). \( \blacksquare \)

C Proof of Lemmas 3, 4, 5, 6, and 7

We first prove the following lemma, which will be useful in subsequent proofs. This lemma does not require that \( f \) be drawn from \( \mathcal{M}_n^{t,k} \).

Lemma 25 Any monotone DNF \( f \) with \( t \geq 2 \) terms each of size \( k \) has \( \Pr[\mathcal{I}] \geq \alpha^3 \).

Proof: We write \( T_1, T_2, \ldots, T_t \) to denote the terms of \( f \). We have

\[
\Pr[\mathcal{I}] = \Pr[T_1 \land T_2 \land \cdots \land T_t] = \Pr[T_1 \mid T_2 \land \cdots \land T_t] \Pr[T_2 \mid T_3 \land \cdots \land T_t] \cdots \Pr[T_{t-1} \mid T_t] \Pr[T_t]
\]

\[
\geq \prod_{i=1}^{t} \Pr[T_i] = \left( 1 - \frac{1}{2^k} \right)^t \geq \left( 1 - \frac{1}{2^k} \right)^{2^{k+1/2} \ln(2/\alpha)} \geq \left( \frac{1}{4} \right)^{2 \ln \frac{2}{\alpha}} > \alpha^3.
\]

The first inequality holds since \( \Pr[f(x) = 1 \mid g(x) = 1] \geq \Pr[f(x) = 1] \) for any monotone Boolean functions \( f, g \) on \( \{0,1\}^n \) (see e.g. Corollary 7, p. 149 of [8]). The second inequality holds by Lemma 2, and the third holds since \( (1 - 1/x)^x \geq 1/4 \) for all \( x \geq 2 \). \( \blacksquare \)

Proof of Lemma 3:

Fix any variable \( v_j \). For each term \( T_k \) we have that \( v_j \) occurs in \( T_k \) with probability \( k/n \). Since the terms are chosen independently, the number of occurrences of \( v_j \) is binomially distributed
according to $B(t, p)$ with $p = k/n$. Taking $\beta = n2^{k-1} \alpha^2/kt^{3/2}\log t$ in the Chernoff bound (which is greater than 1 for sufficiently large $n$), the probability that $v_j$ appears in $\beta pt = 2^{k-1} \alpha^2/\sqrt{t} \log t$ or more terms is at most $\left(\frac{ek^{3/2} \log t}{n2^{k-1} \alpha^2}\right)^{2^{k-1} \alpha^2/\sqrt{t} \log t}$. The lemma follows by the union bound over the $n$ variables $v_j$. \hfill \Box

**Proof of Lemma 4:**

We will modify the proof of the previous lemma slightly. We first fix a value $1 \leq i \leq t$ which will act as the index of a distinguished term $T_i$, and we also fix a value $1 \leq j \leq k$ which will be the index of a distinguished variable within $T_i$. By taking $\beta = \frac{n2^k}{k(t-1) \log t}$ in the Chernoff bound we have that the probability over the choice of the $t - 1$ terms other than $T_i$ that $v_j$ also appears in $\beta p(t-1) = \frac{2^k}{\log t}$ or more terms is at most $\left(\frac{ek(k-1) \log t}{n2^k}\right)^{2^k/\log t}$. We then again apply the union bound, this time over $tk$ different choices of $i$ and $j$. \hfill \Box

**Proof of Lemma 5:**

We are interested in upper bounding the probability $p_i$ that $\log \log t$ or more of the variables in a fixed term $T_i$ belonging to $f$ also appear in some other term $T'_i$ of $f$, for any $\ell \neq i$. First, a simple counting argument shows that the probability that a fixed set of $\log \log t$ variables appears in a set of $k$ variables randomly chosen from among $n$ variables is at most $(k/n)^{\log \log t}$. Since there are $\binom{k}{\log \log t}$ ways to choose a fixed set of $\log \log t$ variables from term $T_i$, we have $p_i \leq \left(\frac{k}{n}\right)^{\log \log t}(t-1)$. The lemma follows by the union bound over the $t$ probabilities $p_i$. \hfill \Box

**Proof of Lemma 6:**

Given an $f$ drawn according to $M_{n,k}$ and given any term $T_i$ in $f$, we are interested in the probability over uniformly drawn instances that $T_i$ is satisfied and $T'_i$ is not satisfied for all $\ell \neq i$. Let $\overline{T_i}$ represent the formula that is satisfied by an assignment $x$ if and only if all of the $T_\ell$ with $\ell \neq i$ are not satisfied by $x$. We want a lower bound on

$$
\Pr[T_i \land \overline{T_i}] = \Pr[\overline{T_i} \mid T_i] \cdot \Pr[T_i].
$$

Since $\Pr[T_i] = 1/2^k$, what remains is to show that with very high probability over random draw of $f$, $\Pr[\overline{T_i} \mid T_i]$ is bounded below by $\alpha^{3/4}$ for all $T_i$. That is, we need to show that $\Pr[\overline{T_i}] \geq \alpha^{3/4}$ with very high probability.

We have that all of the following statements hold with probability at least $1 - \delta_{\text{usat}}$ for every $1 \leq i \leq n$ for a random $f$ from $M_{n,k}$:

1. $\Pr[\overline{T_i}] \geq \prod_{\ell \neq i} \Pr[\overline{T_{\ell,i}}]$: this follows from Equation (*) in the proof of Lemma 25.

2. $\prod_{\ell \neq \ell} \Pr[\overline{T_{\ell,i}}] > \alpha^3$. This holds because the terms in this product are a subset of the terms in Equation (*) in the proof of Lemma 25.

3. At most $2^k/\log t$ terms $T_\ell$ with $\ell \neq i$ are smaller in $f^i$ than they are in $f$ (by Lemma 4).

4. No term in $f^i$ has fewer than $k - \log \log t$ variables (by Lemma 5).

These conditions together imply that $\Pr[\overline{T_i}] \geq \alpha^3 \left(1 - \frac{\log t}{2^k}\right)^{2^k/\log t} \geq \alpha^{3/4}$ using the fact that $(1 - \frac{1}{x})^x \geq 1/4$ for all $x \geq 2$. \hfill \Box

**Proof of Lemma 7:**

By Lemma 5, with probability at least $1 - \delta_{\text{shared}}$ $f$ is such that, for all $1 \leq i < j \leq n$, terms $T_i$ and $T_j$ share at most $\log \log t$ variables. Thus for each pair of terms a specific set of at least $2k - \log \log t$ variables must be simultaneously set to 1 in an instance in order for both terms to be satisfied. \hfill \Box
D Proof of Lemma 8:
P gets a net contribution of 0 from those \( x \) which belong to \( g_{*,*} \) (since each such \( x \) is added twice and subtracted twice in \( P \)). We proceed to analyze the contributions to \( P \) from the remaining 8 subsets of the events \( g_{11}, g_{1*}, \) and \( g_{*1} \):

- \( P \) gets a net contribution of 0 from those \( x \) which are in \( g_{1*} \cap \overline{g}_{*1} \cap \overline{g}_{**} \) since each such \( x \) is counted in \( p_{11} \) and \( p_{10} \) but not in \( p_{01} \) or \( p_{00} \). Similarly \( P \) gets a net contribution of 0 from those \( x \) which are in \( g_{*1} \cap \overline{g}_{*1} \cap \overline{g}_{**} \).

- \( P \) gets a net contribution of \( \Pr [g_{11} \cap \overline{g}_{1*} \cap \overline{g}_{*1} \cap \overline{g}_{**}] \) since each such \( x \) is counted in \( p_{11} \).

- \( P \) gets a net contribution of \( -\Pr [g_{1*} \cap g_{*1} \cap \overline{g}_{**}] \) since each such \( x \) is counted in \( p_{01}, p_{10} \) and \( p_{11} \).

E Proof of Lemma 11
For any fixed \( r \in \{1, \ldots, t\} \) we have that \( \Pr [v_i \text{ and } v_j \text{ cooccur in term } T_i] = \frac{\binom{k}{r} \cdot \binom{k}{r} \cdot \binom{n}{r}}{n^{2r}} \leq \frac{k^2}{n^r} \). Since these events are independent for all \( r \), the probability that there is any collection of \( C \) terms such that \( v_i \text{ and } v_j \text{ cooccur in all } C \) of these terms is at most \( (\frac{1}{C}) \cdot (\frac{k^2}{n^r})^C \leq (\frac{k^2}{n})^C \).

F Proof of Lemma 16
We show that \( P_1 \geq \frac{e^{-2k}}{2^{2k}} \) with probability at least \( 1 - \delta \) if \( P_1 \) is the largest deviation of \( P_1 \) from its expected value is low. Given any fixed length-\( k \) term \( T_1 \), let \( \Omega \) denote the set of all length-\( k \) terms \( T \) which satisfy \( \Pr [T \land T_1] \leq \frac{1}{2^{2k}} \). By reasoning as in the proof of Lemma 15, with probability at least \( 1 - (t - 1)(\frac{k^2}{n})^{log \log t} \) each of \( T_2, \ldots, T_t \) belongs to \( \Omega \), so we henceforth assume that this is in fact the case, i.e. we condition on the event \( \{T_2, \ldots, T_t\} \subset \Omega \). Note that under this conditioning we have that each of \( T_2, \ldots, T_t \) is selected uniformly and independently from \( \Omega \).

We now use McDiarmid’s inequality where the random variables are the randomly selected terms \( T_2, \ldots, T_t \) from \( \Omega \) and \( F(T_2, \ldots, T_t) \) denotes \( P_1 \), i.e.

\[
F(T_2, \ldots, T_t) = \Pr [T_1 \text{ is satisfied by } x \text{ but no } T_j \text{ with } j \geq 2 \text{ is satisfied by } x].
\]

Since each \( T_j \) belongs to \( \Omega \), we have \( |F(T_2, \ldots, T_t) - F(T_2, \ldots, T_{j-1}, T_j', T_{j+1}, \ldots, T_t)| \leq c_t = \frac{\log t}{2^{2k}} \) for all \( j = 2, \ldots, t \). Taking \( \tau = \frac{e^{-2k}}{2^{2k}} \), McDiarmid’s inequality implies that \( \Pr [P_1 \geq \frac{e^{-2k}}{2^{2k}}] \) is at most

\[
\exp \left( \frac{-\alpha^2}{2t} \cdot \frac{\log t}{2^{2k}} \right) = \exp \left( \frac{-\alpha^2}{2t \log^2 t} \right) \leq \exp \left( \frac{-\alpha^2}{2t \log^2 t} \right) \leq \exp \left( \frac{-\alpha^2}{2t \log^2 t} \right)
\]

where the last inequality holds since \( (k, t) \) is \( \alpha \)-interesting. Combining all the failure probabilities, the lemma is proved.

G Proof of Lemmas 17 and 18
Proof of Lemma 17: Since all four of \( g_{00}, g_{01}, g_{10} \) and \( g_{11} \) are empty we need only consider the five events \( g_{**, g_{00}, g_{01}, g_{11}}, g_{1*} \) and \( g_{*1} \). We now analyze the contribution to \( P \) from each possible subset of these 5 events:
• $P$ gets a net contribution of 0 from those $x$ which belong to $g_{s,s}$ (and to any other subset of the remaining four events) since each such $x$ is counted in each of $p_{00}$, $p_{01}$, $p_{10}$ and $p_{11}$. It remains to consider all 16 subsets of the four events $g_{s0}, g_{0s}, g_{s1}$ and $g_{1s}$.

• $P$ gets a net contribution of 0 from those $x$ which are in at least 3 of the four events $g_{s0}, g_{0s}, g_{s1}$ and $g_{1s}$ since each such $x$ is counted in each of $p_{00}$, $p_{01}$, $p_{10}$ and $p_{11}$. $P$ also gets a net contribution of 0 from those $x$ which are in exactly one of the four events $g_{s0}, g_{0s}, g_{s1}$ and $g_{1s}$. It remains to consider those $x$ which are in exactly two of the four events $g_{s0}, g_{0s}, g_{s1}$ and $g_{1s}$.

• $P$ gets a net contribution of 0 from those $x$ which are in $g_{s1}$ and $g_{0s}$ and no other events, since each such $x$ is counted in each of $p_{00}$, $p_{01}$, $p_{10}$ and $p_{11}$. The same is true for those $x$ which are in $g_{s1}$ and $g_{0s}$ and no other events.

• $P$ gets a net contribution of $-\Pr[g_{s1} \land g_{s1} \land (\text{ no other } g_{i} \text{ occurs})]$ from those $x$ which are in $g_{s1}$ and $g_{s1}$ and no other event. Similarly, $P$ gets a net contribution of $-\Pr[g_{0s} \land (\text{ no other } g_{i} \text{ occurs})]$ from those $x$ which are in $g_{0s}$ and $g_{0s}$ and no other event. $P$ gets a net contribution of $\Pr[g_{s1} \land g_{s0} \land (\text{ no other } g_{i} \text{ occurs})]$ from those $x$ which are in $g_{s1}$ and $g_{s0}$ and no other event, and gets a net contribution of $\Pr[g_{0s} \land g_{s1} \land (\text{ no other } g_{i} \text{ occurs})]$ from those $x$ which are in $g_{0s}$ and $g_{s1}$ and no other event.

Proof of Lemma 18: We suppose that $g_{11}$ is nonempty. We wish to analyze the contribution to $P$ from all 64 subsets of the six events $g_{s0}, g_{1s}, g_{0s}, g_{s1}, g_{s0}$ and $g_{11}$. From Lemma 17 we know this contribution for the 32 subsets which do not include $g_{11}$ is (1) so only a few cases remain:

• $P$ gets a net contribution of 0 from those $x$ which are in $g_{11}$ and in $g_{s0}$ and in any other subset of events (each such $x$ is counted in each of $p_{11}, p_{01}, p_{10}$ and $p_{00}$). Similarly, $P$ gets a contribution of 0 from those $x$ which are in $g_{11}$ and in at least three of $g_{s1}, g_{0s}, g_{s1}, g_{s0}$. So it remains only to analyze the contribution from subsets which contain $g_{11}$, contain at most two of $g_{s1}, g_{0s}, g_{s1}, g_{s0}$, and contain nothing else.

• An analysis similar to that of Lemma 17 shows that $P$ gets a net contribution of $\Pr[g_{11} \land g_{s1} \land g_{s0} \land (\text{ no other } g_{i} \text{ occurs})] + \Pr[g_{11} \land g_{s0} \land g_{s1} \land (\text{ no other } g_{i} \text{ occurs})] - \Pr[g_{11} \land g_{s1} \land g_{s0} \land (\text{ no other } g_{i} \text{ occurs})] + \Pr[g_{11} \land g_{s1} \land g_{s0} \land (\text{ no other } g_{i} \text{ occurs})]$ from those $x$ which are in $g_{11}$, in exactly two of $g_{s1}, g_{0s}, g_{s1}, g_{s0}$, and in no other events. So it remains only to consider subsets which contain $g_{11}$ and at most one of $g_{s1}, g_{0s}, g_{s1}, g_{s0}$ and nothing else.

• $P$ gets a contribution of 0 from those $x$ which are in $g_{11}$ and in $g_{s1}$ and in nothing else; likewise from $x$ which are in $g_{11}$ and in $g_{s1}$ and in nothing else. $P$ gets a contribution of $\Pr[g_{11} \land g_{s0} \land (\text{ no other } g_{i} \text{ occurs})]$ from those $x$ which are in $g_{11}$ and in $g_{0s}$ and in nothing else, and a contribution of $\Pr[g_{11} \land g_{s0} \land (\text{ no other } g_{i} \text{ occurs})]$ from those $x$ which are in $g_{11}$ and in $g_{s0}$ and in nothing else.

• $P$ gets a net contribution of $\Pr[g_{11} \land (\text{ no other } g_{i} \text{ occurs})]$ from those $x$ which are in $g_{11}$ and in no other event.

H Proof of Lemma 20

We use McDiarmid’s inequality, where the random variables are the terms $T_1, \ldots, T_t$ chosen independently from the set of all possible terms of length $k$ and $F(T_1, \ldots, T_t)$ denotes the number of pairs of variables $(v_i, v_j)$ that cooccur in at least two terms. For each $\ell = 1, \ldots, t$ we have $\Pr[T_\ell \text{ contains both } v_1 \text{ and } v_2] \leq \frac{k^2}{n^2}$, so by a union bound we have $\Pr[f \text{ contains at least two terms which contain both } v_1 \text{ and } v_2] \leq \frac{t^2k^4}{n^2}$. By linearity of expectation
we have \( \mu = \operatorname{E}[F] \leq \frac{2^k t}{n^k} \). Since each term involves at most \( k^2 \) pairs of cooccurring variables, we have \( |F(T_1, \ldots, T_i) - F(T_1, \ldots, T_{i-1}, T_i, T_{i+1}, \ldots, T_t)| \leq \epsilon_i = k^2 \). We thus have by McDiarmid’s inequality that \( \Pr[F \geq \frac{k^2 t^4}{n^2} + \tau] \leq \exp(-\tau^2/(tk^4)) \). Taking \( \tau = dt^2 k^4/n^2 \), we have \( \Pr[F \geq (d + 1)t^2 k^4/n^2] \leq \exp(-dt^2 k^4/n^2) \).

\[ \]

**I Proof of Lemma 22**

In order for \( v_1, v_2, v_j \) to form a triangle in \( G \), it must be the case that either (i) some clique \( S_i \) contains \( \{1, 2, j\} \); or (ii) there is some pair of cliques \( S_a, S_b \) with \( 2 \notin S_a \) and \( \{1, j\} \subset S_a \) and \( 1 \notin S_b \) and \( \{2, j\} \subset S_b \).

For (i), we have from Lemma 11 that \( v_1 \) and \( v_2 \) cooccur in more than \( C \) terms with probability at most \( \left( \frac{\binom{k^2 + 1}{n}^2}{n^2} \right)^C \). Since each term in which \( v_1 \) and \( v_2 \) cooccur contributes at most \( k - 2 \) vertices \( v_j \) to condition (i), the probability that more than \( C(k - 2) \) vertices \( v_j \) satisfy condition (i) is at most \( \left( \frac{\binom{k^2 + 1}{n}^2}{n^2} \right)^C = O(1/n^{C/2}). \)

For (ii), let \( A \) be the set of those indices \( a \in \{1, \ldots, t\} \) such that \( 2 \notin S_a \) and \( 1 \in S_a \), and let \( S_A \) be \( \cup_{a \in A} S_a. \) Similarly let \( B \) be the set of indices \( b \) such that \( 1 \notin S_b \) and \( 2 \in S_b \), and let \( S_B \) be \( \cup_{b \in B} S_b. \) It is clear that \( A \) and \( B \) are disjoint. For each \( \ell = 1, \ldots, t \) we have that \( \ell \in A \) independently with probability at most \( p = \frac{k}{n} \), so \( E[|A|] \leq \frac{kt}{n} \). We now consider two cases:

**Case 1: \( t \leq n/\log n \)**. In this case we may take \( \beta = \frac{\log n}{tk} \) in the Chernoff bound, and we have that \( \Pr[|A| \geq \beta pt] \) equals

\[
\Pr[|A| \geq \beta pt] = \Pr[|A| \geq \frac{tk \log n}{n}] \leq \left( \frac{e}{\log n} \right)^{\frac{tk \log n}{n}} \leq \left( \frac{e}{\log n} \right)^{\frac{kt \log n}{n}} = \frac{1}{n^{\omega(1)}}.
\]

The same bound clearly holds for \( B \). Note that in Case 1 we thus have \( |S_A|, |S_B| \leq k \log n \) with probability \( 1 - 1/n^{\omega(1)} \).

**Case 2: \( t > n/\log n \)**. In this case we may take \( \beta = \log n \) in the Chernoff bound and we obtain

\[
\Pr[|A| \geq \beta pt] = \Pr[|A| \geq \frac{tk \log n}{n}] \leq \left( \frac{e}{\log n} \right)^{\frac{tk \log n}{n}} \leq \left( \frac{e}{\log n} \right)^{\frac{k \log n}{n}} = \frac{1}{n^{\omega(1)}}
\]

where the last inequality holds since \( k = \Omega(\log n) \) (since \( t > n/\log n \) and \( (k, t) \) is \( \alpha \)-interesting). In Case 2 we thus have \( |S_A|, |S_B| \leq \frac{tk \log n}{n} \) with probability \( 1 - 1/n^{\omega(1)} \).

Let \( S'_A \) denote \( S_A - \{1\} \) and \( S'_B \) denote \( S_B - \{2\} \). Since \( A \) and \( B \) are disjoint, it is easily seen that conditioned on \( S'_A \) being of some particular size \( s'_A \), all \( \binom{n-2}{s'_A} \) \( s'_A \)-element subsets of \( \{3, \ldots, n\} \) are equally likely for \( S'_A \). Likewise, conditioned on \( S'_B \) being of size \( s'_B \), all \( \binom{n-2}{s'_B} \) \( s'_B \)-element subsets of \( \{3, \ldots, n\} \) are equally likely for \( S'_B \). Thus, the probability that \( |S'_A \cap S'_B| \geq C \) is at most

\[
\binom{s'_B}{C} \left( \frac{s'_A}{n-2} \right)^C \leq \frac{s'_A^C}{s'_B^C} \leq \left( \frac{s'_A}{n-2} \right)^C.
\]

(since the expression on the left is an upper bound on the probability that any collection of \( C \) elements in \( S'_B \) all coincide with elements of \( S'_A \)).

In Case 1 (\( t \leq n/\log n \)) we may assume that \( s'_A, s'_B \) are each at most \( k \log n \), and thus (2) is at most \( C(n/\log n)^C \). In Case 2 (\( t > n/\log n \)) we may assume that \( s'_A, s'_B \leq \frac{tk \log n}{n} \), and thus (2) is at most \( \frac{2^k n^2 \log^2 n}{n^2} \). Thus all in all, we have that except with probability \( O(1/n^{C/2}) \) event (i) contributes at most \( C(k - 2) \) vertices \( v_j \) such that \( \{1, 2, j\} \) forms a triangle, and except with probability \( O(\log^{6C} n) \) event (ii) contributes at most \( C \) vertices \( v_j \) such that \( \{1, 2, j\} \) forms a triangle. This proves the lemma.