DNF are Teachable in the Average Case

Homin K. Lee, Rocco A. Servedio*, and Andrew Wan**

Columbia University, New York, NY 10027, USA

Abstract. We study the average number of well-chosen labeled examples that are required for a helpful teacher to uniquely specify a target function within a concept class. This "average teaching dimension" has been studied in learning theory and combinatorics and is an attractive alternative to the "worst-case" teaching dimension of Goldman and Kearns [7] which is exponential for many interesting concept classes. Recently Balbach [3] showed that the classes of 1-decision lists and 2-term DNF each have linear average teaching dimension.

As our main result, we extend Balbach's teaching result for 2-term DNF by showing that for any $1 \le s \le 2^{\Theta(n)}$, the well-studied concept classes of at-most-s-term DNF and at-most-s-term monotone DNF each have average teaching dimension $O(ns)$. The proofs use detailed analyses of the combinatorial structure of "most" DNF formulas and monotone DNF formulas. We also establish asymptotic separations between the worst-case and average teaching dimension for various other interesting Boolean concept classes such as juntas and sparse GF_2 polynomials.

1 Introduction

Many results in computational learning theory consider learners that have some form of access to an oracle that provides labeled examples. Viewed as teachers, these oracles tend to be unhelpful as they typically either provide random examples selected according to some distribution, or they put the onus on the learner to select the examples herself. In noisy learning models, oracles are even allowed to lie from time to time.

In this paper we study a learning model in which the oracle acts as a helpful teacher [7, 8]. Given a target concept c (this is simply a Boolean function over some domain X) that belongs to a concept class $\mathcal{C}$, the teacher provides the learner with a carefully chosen set of examples that are labeled according to c. This set of labeled examples is called a *teaching set* and must have the property that no other concept $c' \neq c$ in $\mathcal{C}$ is consistent with the teaching set; thus every learner that outputs a consistent hypothesis will correctly identify c as the target concept. The minimum number of examples in any teaching set for c is called the *teaching dimension of c with respect to $\mathcal{C}$*, and the maximum value of the teaching dimension over all concepts in $\mathcal{C}$ is the *teaching dimension of $\mathcal{C}$*.

* Supported in part by NSF award CCF-0347282, by NSF award CCF-0523664, and by a Sloan Foundation Fellowship.

** H. K. Lee and A. Wan are eligible for Best Student Paper.

Some concept classes that are easy to learn can be very difficult to teach in the worst case in this framework. As one example, let the concept class $\mathcal{C}$ over finite domain X contain $|X|+1$ concepts which are the $|X|$ singletons and the empty set. Any teaching set for the empty set must contain every example in X, since if $x \in X$ is missing from the set then the singleton concept $\{x\}$ is not ruled out by the set. Thus the teaching dimension for this concept class is $|X|$.

Many interesting concept classes include the empty set and all singletons, and thus have teaching dimension $|X|$. Consequently for many concept classes the (worst-case) teaching dimension is not a very interesting measure. With this motivation, researchers have considered the *average teaching dimension*, namely the average value of the teaching dimension of c as c ranges over all of $\mathcal{C}$.

Anthony *et al.* [2] showed that the average teaching dimension of the class of linearly separable Boolean functions over $\{0,1\}^n$ is $O(n^2)$. Kuhlmann [9] showed that concept classes with VC dimension 1 over finite domains have constant average teaching dimension and also gave a bound on the average teaching dimension of concept classes $\mathcal{B}^d(c)$ (balls of center c and size $\leq d$). Kushilevitz *et al.* [10] constructed a concept class $\mathcal{C}$ that has an average teaching dimension of $\Omega(\sqrt{|\mathcal{C}|})$ (this lower bound was also proved in [6]) and also showed that every concept class has average teaching dimension at most $O(\sqrt{|\mathcal{C}|})$. More recently, Balbach [3] showed that the classes of 2-term DNF and 1-decision lists each have average teaching dimension linear in n.

Our Results. Our main results are the following theorems, proved in Sections 3 and 4, which show that the well-studied concept classes of monotone DNF formulas and DNF formulas are efficiently teachable in the average case:

Theorem 1. *Fix any $1 \leq s \leq 2^{\Theta(n)}$ and let $\mathcal{C}$ be the concept class of all Boolean functions over $\{0,1\}^n$ representable as a monotone DNF with at most s terms. Then the average teaching dimension of $\mathcal{C}$ is $O(ns)$.*

Theorem 2. *Fix any $1 \leq s \leq 2^{\Theta(n)}$ and let $\mathcal{C}$ be the concept class of all Boolean functions over $\{0,1\}^n$ representable as a DNF with at most s terms. Then the average teaching dimension of $\mathcal{C}$ is $O(ns)$.*

Theorem 2 is a broad generalization of Balbach's result on the average teaching dimension of the concept class of DNF with at most two terms. It is easy to see that even the class of at-most-2-term DNFs has exponential worst-case teaching dimension; as we show in Section 3, the worst-case teaching dimension of at-most-s-term monotone DNFs is exponential as well. Thus our results show that there is a dramatic difference between the worst-case and average teaching dimensions for these concept classes.

We also consider some other well-studied concept classes, namely juntas and sparse GF_2 polynomials. For the class of k-juntas, we show in Section 5 that while the worst-case teaching dimension has a logarithmic dependence on n (the number of irrelevant variables), the average teaching dimension has no dependence on n. For a certain class of sparse GF_2 polynomials (roughly, the class of GF_2 polynomials with fewer than $\log n$ terms; see Section 6), we show that while the worst-case teaching dimension is $n^{\Theta(\log\log n)}$, the average teaching dimension is $O(n \log n)$. Thus in each case we establish an asymptotic separation between the worst-case teaching dimension and the average teaching dimension. Our results suggest that rich and interesting concept classes that are difficult to learn in many models may in fact be easy to teach in the average case.

2 Preliminaries

Our domain is $X = \{0,1\}^n$, and we refer to Boolean functions $c : \{0,1\}^n \to \{0,1\}$ as *concepts*. A collection of concepts $\mathcal{C} \subseteq 2^{\{0,1\}^n}$ is a *concept class*. For a given instance $x \in X$, the value of $c(x)$ is referred to as a *label*, and for $y \in \{0,1\}$, the pair (x, y), is referred to as a *labeled example*. If $y = 0$ ($y = 1$) then the pair is called a *negative* (*positive*) *example*. A concept class $\mathcal{C}$ is *consistent* with a set of labeled examples if $c(x) = y$ for all the examples in the set.

A set S of labeled examples is a *teaching set for c with respect to* $\mathcal{C}$ if c is the only concept in $\mathcal{C}$ that is consistent with S; thus every learner that outputs a consistent hypothesis from $\mathcal{C}$ will correctly identify c as the target concept. The minimum number of examples in any teaching set for c is called the *teaching dimension of c with respect to* $\mathcal{C}$ (sometimes written $TD(c)$ when $\mathcal{C}$ is understood), and the maximum value of the teaching dimension over all concepts in $\mathcal{C}$ is the (worst-case) *teaching dimension of* $\mathcal{C}$. The *average teaching dimension* of $\mathcal{C}$ is the average value of the teaching dimension of c with respect to $\mathcal{C}$ for all c, *i.e.*, $\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} TD(c)$.

We use Boolean variables $x_1, \ldots, x_n$ and write $\bar{x}_i$ to denote the negated literal on variable x_i. We will often refer to a logical assignment of the variables as a string and vice-versa; thus, a string $y \in \{0,1\}^n$ corresponds to a truth-value assignment to the variables $x_1, \ldots, x_n$. Given a set S of variables, we write $\mathbf{0}|_{S=1}$ to denote the truth assignment that sets each variable in S to 1 and sets all other variables to 0. The truth assignment $\mathbf{1}|_{S=0}$ is defined similarly.

Two strings $y, z \in \{0,1\}^n$ are *neighbors* if they differ in exactly one bit position. Given $x, y \in \{0,1\}^n$ we write $x \leq y$ if $x_i \leq y_i$ for all $i = 1, \ldots, n$, and we write $x < y$ if we have $x \leq y$ and $x \neq y$.

DNF Formulas. A *term* is a conjunction of Boolean literals. A term over n variables is represented by a string $T \in \{0,1,*\}^n$, where the k-th character of T is denoted $T[k]$. The value of $T[k]$ is 0, 1, or $*$ depending on whether x_k occurs negated, unnegated, or not at all in the term. If $x \in \{0,1\}^n$ is an assignment that satisfies T, we sometimes say that T *covers* x. Note that the satisfying assignments of a term T form a subcube of dimension $n - |T|$ within $\{0,1\}^n$, where $|T|$ denotes the number of non-$*$ entries in T.

An *s-term DNF formula* ϕ is an OR of s terms $\phi = T_1 \vee \cdots \vee T_s$. A satisfying assignment to the DNF is sometimes referred to as a *positive point* and an unsatisfying assignment as a *negative point.*

A term T_i is said to be *compatible* with a set of labeled examples S if T_i does not cover any negative example in S. A term T_i is said to *imply* another term T_j if every positive point of T_i is also a positive point of T_j. We similarly say that a term T implies a DNF formula ϕ, or that a DNF formula ϕ_1 implies another DNF formula ϕ_2. Two different DNF formulas ϕ_1 and ϕ_2 are said to be *logically equivalent* if each implies the other, *i.e.,* if they are different syntactic representations of the same Boolean function. Throughout the paper we will use Greek letters $\phi, \varphi, \ldots$ to denote formulas (which are syntactic objects) and Roman letters $f, g, \ldots$ to denote Boolean functions (which are abstract mappings from $\{0,1\}^n$ to $\{0,1\}$).

We write $\mathcal{D}_s$ to denote the class of "exactly-s-term" DNFs; this is the class of all Boolean functions $f\colon \{0,1\}^n \to \{0,1\}$ that have some s-term DNF representation and have no s'-term DNF representation for any $s' < s$. Similarly, we write $\mathcal{D}_{\leq s}$ to denote the class of "at-most-s-term" DNFs, which is $\mathcal{D}_{\leq s} = \cup_{s' \leq s} \mathcal{D}_{s'}$. Note that the elements of $\mathcal{D}_s$ and $\mathcal{D}_{\leq s}$ are "semantic" functions, not syntactic formulas. The class $\mathcal{D}_{\leq s}$ corresponds to the standard notion of "s-term DNF" which is a well studied concept class in computational learning theory.

A *monotone DNF formula*, or mDNF, is a DNF formula that contains no negated literals. The concept classes of exactly-s-term mDNFs and at-most-s-term mDNFs are denoted $\mathcal{M}_s$ and $\mathcal{M}_{\leq s}$ and are defined in analogy with $\mathcal{D}_s$ and $\mathcal{D}_{\leq s}$ above. The following fact is well known:

Fact 1 *If $f \in \mathcal{M}_s$ then there is a unique (up to ordering of the terms) s-term mDNF representation $\phi = T_1 \vee \cdots \vee T_s$ for f.*

3 Monotone DNFs

Worst-case teaching dimension of at-most-s-term mDNFs. Here we state upper and lower bounds on the worst-case teaching dimension of $\mathcal{M}_{\leq s}$.

Theorem 3. *The teaching dimension of $\mathcal{M}_{\leq s}$ is at most $n^s + s$.*

Proof. Let f be an element of $\mathcal{M}_k$ for some $k \leq s$. We have that f is represented by a unique mDNF $\phi = T_1 \vee \cdots \vee T_k$, where each T_k corresponds to a minterm (minimal satisfying assignment) of f. For the rest of the proof we will view each term T_i as the set of variables that it contains; note that these sets are pairwise incomparable, *i.e.,* no T_i is contained in any other T_j.

We will show that the following set of examples is a teaching set for f:

- For each term T_i in $T_1, \ldots, T_k$ we give the positive example $\mathbf{0}|_{T_i=1}$; this is clearly at most s examples.
- We also give a set of negative examples which consists of precisely those examples that have exactly one variable of each term set to zero and all other variables set to one. In other words, for every set $S \subseteq \cup_{i=1}^{k} T_i$ that satisfies $|S \cap T_i| = 1$ for all i, we give the example $\mathbf{1}|_{S=0}$. Since there are at most n^s ways to choose exactly one variable from each of the s terms, this is at most n^s examples.

We first note that any $g \in \mathcal{M}_{\leq s}$ (in fact any monotone function g) that is consistent with the negative set must label negative any assignment which does not satisfy at least one of the terms $T_1, \ldots, T_k$. This is because for any assignment y which satisfies none of the k terms, there is an example y' in the negative set such that $y \leq y'$ with respect to the bitwise partial order on $\{0,1\}^n$. Since $g(y') = 0$ and g is monotone, this implies that $g(y) = 0$.

It follows that for all T_i, for all $x_j \in T_i$, the example $\mathbf{0}|_{(T_i \setminus x_j)=1}$ must be negative under f. Thus for each term T_i, we have that $\mathbf{0}|_{T_i=1}$ is positive while flipping any bit in T_i makes f negative. Consequently, any $g \in \mathcal{M}_{\leq s}$ which is consistent with both the positive and negative examples must contain each of the terms $T_1, \ldots, T_k$. If $k = s$ then since g cannot contain any other terms, we must have that g is equivalent to f. If $k < s$, suppose that g contains some other non-redundant term T_{s+1}. Then there must be an assignment that is positive under g but which does not satisfy any of $T_1, \ldots, T_k$. The negative set shows that this is not possible. □

Theorem 4. *Given s, let $s' \leq s$ be any value such that $(s'-1)$ divides n. Then the teaching dimension of $\mathcal{M}_{\leq s}$ is at least $(\frac{n}{s'-1})^{s'-1}$.*

Proof. We exhibit a concept $f \in \mathcal{M}_{\leq s-1}$ whose teaching set must contain all the negative examples in the teaching set for the proof of Theorem 3 in order to disambiguate it from various concepts in $\mathcal{M}_s$. Let $d = n/(s'-1)$ and consider the concept

$$f = (x_1 \cdots x_d) \vee (x_{d+1} \cdots x_{2d}) \vee \cdots (x_{(s'-2)d} \cdots x_{(s'-1)d})$$

which is known as the *tribes* function. Suppose not all of the negative examples from the proof of Theorem 3 are part of a teaching set for f, *i.e.,* that there is some $S \subseteq \cup_{i=1}^{s'-1} T_i$ such that $\mathbf{1}|_{S=0} \notin TS(f)$. Let T_S be the term exactly satisfied by $\mathbf{1}|_{S=0}$ so that T_S includes variable x_i if and only if the i'th bit of $\mathbf{1}|_{S=0}$ is set to 1. Then the concept $f' = f \vee T_S$ will label $TS(f)$ consistently with f. Clearly any positive example in $TS(f)$ will also be positive under f'. Take any negative example $y \in TS(f)$. Unless y satisfies T_S, it is negative under f'. But to satisfy T_S, $y \leq \mathbf{1}|_{S=0}$. This is impossible, since if $y < \mathbf{1}|_{S=0}$ it would be a positive example, and $y = \mathbf{1}|_{S=0}$ is not in the teaching set by assumption.

Thus $TS(f)$ must contain a negative example for every $S \subseteq \cup_{i=1}^{s'-1} T_i$ satisfying $|S \cap T_i| = 1$ for all i. For f there are $d^{s'-1} = (\frac{n}{s'-1})^{s'-1}$ such sets. □

Average-case teaching dimension of at-most-s-term mDNFs. We now prove Theorem 1. The idea is to show that almost every at-most-s-term monotone DNF in fact has exactly s terms; as we will see, these exactly-s-term monotone DNFs can be taught very efficiently with $O(ns)$ examples. The remaining concepts are so few that they can be handled with a brute-force approach and the overall average teaching dimension will still be $O(ns)$.

We start with a simple lemma from [7]:

Lemma 1 ([7]). *Let c be any concept in $\mathcal{M}_s$. Then the teaching dimension of c with respect to $\mathcal{M}_{\leq s}$ is at most $(n+1)s$.*

Proof sketch: Let $\phi = T_1 \vee \cdots \vee T_s$ be the unique s-term mDNF for c. For each $i = 1, \ldots, s$ the teaching set contains the positive example $\mathbf{0}|_{T_i=1}$ and contains $|T_i|$ many negative examples which are the neighbors of $\mathbf{0}|_{T_i=1}$ that are obtained by setting one of the 1s to 0. Each of the s terms thus contributes at most $n+1$ examples; an easy argument based on Fact 1 given in [7] shows that this is indeed a teaching set. $\square$

Lemma 2. *For $1 \le i < \frac{1}{4}e^{\frac{n}{72}}$, we have $\frac{2^{ni-1}}{i!} \le |\mathcal{M}_i| \le \frac{2^{ni}}{i!}$.*

Proof. The upper bound is easy: the number of i-term mDNFs is at most the number of ways to choose i terms from the set of all 2^n many monotone terms over variables $x_1, \ldots, x_n$. The latter quantity is $\binom{2^n}{i} \le \frac{2^{ni}}{i!}$.

For the lower bound we consider all 2^{ni} ways to select a sequence of i terms (with replacement) from the set of all 2^n possible monotone terms. We show that at least half of these 2^{ni} ways result in a sequence $T_1, \ldots, T_i$ of terms which are pairwise incomparable, *i.e.,* no T_i implies any other T_j. Each such sequence yields an i-term mDNF, and each such mDNF occurs $i!$ times because of different orderings of the terms in a sequence. This gives the lower bound.

Note that a collection of i monotone terms $T_1, \ldots, T_i$ will be pairwise incomparable if the following two conditions hold: (1) Each of the i terms contains between $5n/12$ and $7n/12$ many variables, and (2) Viewing each term T_i as a set of variables, for any $j \ne k$ the symmetric difference $|T_j \Delta T_k|$ is of size at least $n/4$. (This is because if $|T_j|, |T_k| \in [5n/12, 7n/12]$ and $T_j \subseteq T_k$, then the symmetric difference must be of size at most $n/6$.)

For condition (1), Hoeffding's bound implies that a uniformly selected monotone term T will contain fewer than $5n/12$ or more than $7n/12$ many variables with probability at most $2e^{-n/72}$, so a union bound gives that condition (1) fails with probability at most $2ie^{-n/72}$. For condition 2, observe that given two uniform random terms T_j, T_k, each variable x_ℓ is independently in their symmetric difference with probability 1/2. Thus Hoeffding's bound implies that $|T_j \Delta T_k| < n/4$ with probability at most $e^{-n/8}$. By a union bound, the probability that condition (2) fails is at most $\binom{i}{2}e^{-n/8}$. Thus for $i < \frac{1}{4}e^{\frac{n}{72}}$, the probability that conditions (1) and (2) both hold is at least 1/2. $\square$

Fix $1 \le s \le \frac{1}{4}e^{\frac{n}{72}}$. It is easy to check that by Lemma 2, for any $k < s$ we have $|\mathcal{M}_k| < \frac{1}{2}|\mathcal{M}_{k+1}|$. Thus (again by Lemma 2) we have $|\mathcal{M}_{\le s-1}| \le \frac{2^{ns-n+1}}{(s-1)!}$ while $|\mathcal{M}_s| \ge \frac{2^{ns-1}}{s!}$. Combining these bounds gives that $\frac{|\mathcal{M}_s|}{|\mathcal{M}_{\le s-1}|} \ge \frac{2^n}{4s}$. By Lemma 1, each concept $c \in \mathcal{M}_{\le s}$ which is in $\mathcal{M}_s$ can be taught using $n(s+1)$ examples. Each of the remaining concepts can surely be taught using at most 2^n examples. We thus have that the average teaching dimension of $\mathcal{M}_{\le s}$ is at most

$$\frac{(n+1)s|\mathcal{M}_s| + 2^n|\mathcal{M}_{\le s-1}|}{|\mathcal{M}_s| + |\mathcal{M}_{\le s-1}|} \le (n+1)s + \frac{2^n}{1 + 2^n/4s} \le (n+1)s + 4s,$$

giving us the following result which is a slightly sharper version of Theorem 1:

Theorem 5. *Let s be any value $1 \le s \le \frac{1}{4}e^{\frac{n}{72}}$. The class $\mathcal{M}_{\le s}$ of at-most-s-term monotone DNF has average teaching dimension at most $s(n+5)$.*

Note that if $s > \frac{1}{4}e^{\frac{n}{72}}$, then 2^n is bounded by some fixed polynomial in s, and thus the worst-case teaching number 2^n is actually $\mathrm{poly}(n, s)$ for such a large s. This gives the following corollary which says that the class of at-most-s-term monotone DNF is efficiently teachable on average for all possible values of s:

Corollary 1. *Let s be any value $1 \le s \le 2^n$. The class $\mathcal{M}_{\le s}$ of at-most-s-term monotone DNF has average teaching dimension poly(n, s).*

4 DNFs

Now we will tackle the teaching dimension of the unrestricted class of size-at-most-s DNFs. The high-level approach is similar to the monotone case, but the details are more complicated. The idea is to identify a subset $\mathcal{S}$ of $\mathcal{D}_{\le s}$ and show that (i) any function $f \in \mathcal{S}$ can be uniquely specified within all of $\mathcal{D}_{\le s}$ using only $O(ns)$ examples; and (ii) at most a $\frac{O(s)}{2^n}$ fraction of all functions in $\mathcal{D}_{\le s}$ do not belong to $\mathcal{S}$. Given (i) and (ii) it is easy to conclude that the average teaching number of $\mathcal{D}_{\le s}$ is $O(ns)$.

The challenge is to devise a set $\mathcal{S}$ that satisfies both conditions (i) and (ii). In the monotone case using Fact 1 it was easy to show that $\mathcal{M}_s$ is an easy-to-teach subset, but non-monotone DNF are much more complicated (no analogue of Fact 1 holds for non-monotone DNF) and it is not at all clear that all functions in $\mathcal{D}_s$ are easy to teach. Thus we must use a more complicated set $\mathcal{S}$ of easy-to-teach functions; we define this set and prove that it is indeed easy to teach in Section 4.2. (This argument uses Balbach's results for exactly-2-term DNFs.) The argument that (ii) holds for $\mathcal{S}$ is correspondingly more complex than the counting argument for mDNFs because of $\mathcal{S}$'s more involved structure; we give this in Section 4.3.

4.1 Preliminaries

We will borrow some terminology from Balbach [3]. Two terms T_i and T_j have a *strong difference* at k if $T_i[k], T_j[k] \in \{0, 1\}$ and $T_i[k] \ne T_j[k]$ (*e.g.,* $x_1\bar{x}_5x_6$ and $\bar{x}_5\bar{x}_6x_{12}x_{23}$ have a strong difference at position 6). Two terms have a *weak difference* at k if $T_i[k] \in \{0, 1\}$ and $T_j[k] = *$ or vice-versa. Two weak differences at positions k and ℓ are *of the same kind* if $T_i[k], T_i[\ell] \in \{0, 1\}$ and $T_j[k] = T_j[\ell] = *$ or vice-versa, that is both $*$'s occur in the same term (*e.g.,* $\bar{x}_5x_6$ and $\bar{x}_5\bar{x}_6x_{12}x_{23}$ have two weak differences of the same kind at positions 12 and 23). Two weak differences at positions k and ℓ are *of different kinds* if $T_i[k], T_j[\ell] \in \{0, 1\}$ and $T_j[k] = T_i[\ell] = *$ or vice-versa (*e.g.,* $\bar{x}_5x_6$ and $\bar{x}_5x_{12}$ have two weak differences of different kinds at positions 6 and 12).

Now we introduce some new terminology. Given $y \in \{0, 1\}^n$ which satisfies a term T, we denote by $N_T(y)$ the set consisting of y and all its neighbors that do *not* satisfy T. A satisfying assignment $y \in \{0, 1\}^n$ of a term T in ϕ is called a *cogent corner point of* T if all the neighbors of y that satisfy ϕ satisfy T, and all the neighbors that do not satisfy T do not satisfy ϕ. Note that if y is a cogent corner point of T, then each of the neighbors of y in $N_T(y)$ does not satisfy ϕ. A pair of points $y, z \in \{0,1\}^n$ that satisfy a term T are said to be *antipodal around* T if $y_k = \overline{z}_k$ for all k such that $T[k] = *$. A pair of points are *cogent antipodal points around* T if they are both cogent corner points of T and antipodal around T. This leads us to our first preliminary lemma:

Lemma 3. *Let $\phi = T_1 \vee \cdots \vee T_s$ be any DNF. Let y be a cogent corner point of T_i. Any $\widehat{T}$ that covers y and is compatible with $N_{T_i}(y)$ must imply T_i.*

Proof. Let $\widehat{T}$ be any term that covers y. Observe that for each literal ℓ in T_i, if $\widehat{T}$ did not contain ℓ then $\widehat{T}$ would not be compatible with $N_{T_i}(y)$ since the corresponding negative neighbor of y is contained in $N_{T_i}(y)$ but would be covered by $\widehat{T}$. It follows that every literal in T_i is also present in $\widehat{T}$, and consequently $\widehat{T}$ implies T_i. □

Two terms are said to be *close* if they have at most one strong difference. Note that there is no strong difference between two terms if and only if they have some satisfying assignment in common, and there is one strong difference between two terms if and only if they have neighboring satisfying assignments.

Given a Boolean function $f \colon \{0,1\}^n \to \{0,1\}$, we let G_f denote the undirected graph whose vertices are the satisfying assignments of f and whose edges are pairs of neighboring satisfying assignments. A *cluster* C of f is a set of satisfying assignments that form a connected component in G_f. We sometimes abuse notation and write C to refer to the Boolean function whose satisfying assignments are precisely the points in C. We say that a DNF ϕ computes cluster C if the set of satisfying assignments for ϕ is precisely C. The *DNF-size* of a cluster C is the minimum number of terms in any DNF that computes C. For intuition, we can view a cluster as being a connected set of positive points that have a "buffer" of negative points separating them from all other positive points. The following lemma is immediate:

Lemma 4. *Let f be an element of $\mathcal{D}_s$, i.e. f is an exactly-s-term DNF. Let $C_1, \ldots, C_r$ be the clusters of f. Then DNF-size$(C_1) + \cdots +$ DNF-size$(C_r) = s$.*

4.2 Teaching $\mathcal{S}$

We are now ready to define our "nice" (easy to teach) subset $\mathcal{S} \subseteq \mathcal{D}_{\leq s}$ of size-at-most-s DNFs. (We emphasize that $\mathcal{S}$ is a set of functions, not of DNF expressions.) $\mathcal{S}$ consists of those exactly-s-term DNFs (so in fact $\mathcal{S} \subseteq \mathcal{D}_s$) all of whose clusters either: (1) have DNF-size 1; (2) have DNF-size 2; or (3) have DNF-size k, for some k, and are computed by a DNF $\phi = T_1 \vee \cdots \vee T_k$ in which each T_i has a pair of cogent antipodal points around it.

Note that if a cluster has DNF-size 1, then it clearly satisfies condition (3) above (in fact every pair of antipodal points for the term is cogent). Thus we can simplify the description of $\mathcal{S}$: it is the set of all exactly s-term DNFs all of whose clusters either: (i) have DNF-size k and are computed by a DNF $\phi = T_1 \vee \cdots \vee T_k$ in which each T_i has a pair of cogent antipodal points around it, or (ii) have

DNF-size exactly 2. (Note that there do in fact exist Boolean functions of DNF-size 2 for which any two-term representation $T_1 \vee T_2$ has some term T_i with no pair of cogent antipodal points around it, *e.g.,* $\overline{x}_1\overline{x}_3 \vee x_2x_3$, and thus condition (ii) is non-redundant.)

The teaching set for functions in $\mathcal{S}$. We will use the following theorem due to Balbach [3]:

Theorem 6. *Let c be any element of $\mathcal{D}_2$ (i.e., an exactly-2-term DNF). The teaching dimension of c with respect to $\mathcal{D}_{\leq 2}$ is at most $2n+4$.*

The teaching set specified in [3] to prove Theorem 6 consists of at most 5 positive points along with some negative points. Given $f \in \mathcal{D}_2$, we define $BTS(f)$ to be the union of the teaching set specified in [3] together with all negative neighbors of the (at most five) positive points described above (the set specified in [3] already contains some of these points). With this definition a straightforward consequence of the analysis of [3] is the following:

Lemma 5. *Let $\phi = T_1 \vee \cdots \vee T_s$ be a DNF that has a cluster C with DNF-size 2. Let $BTS(C)$ be as described above. Let y be a satisfying assignment for ϕ that is contained in C. Then any term $\widehat{T}$ that covers y and is consistent with $BTS(C)$ must imply C.*

Given any function $f \in \mathcal{S}$, our teaching set $TS(f)$ for f will be as follows. For each cluster C of f, if C:

- **satisfies condition (i):** then for each term T_i described in condition (i), the set $TS(f)$ contains a pair y, z of cogent antipodal points for T_i (these are positive examples) and contains all negative neighbors of these two positive examples (*i.e.,* $TS(f)$ contains $N_{T_i}(y)$ and $N_{T_i}(z)$). Thus $TS(f)$ includes at most $k(2+2n)$ many points from such a cluster.
- **does not satisfy condition (i) but satisfies (ii):** then we will give the set $BTS(C)$ described above. By Theorem 6 and the definition of $BTS(C)$, we have that $BTS(C)$ contains at most $7n+4$ points.

Lemma 4 now implies that $TS(f)$ contains at most $O(ns)$ points.

Correctness of the teaching set construction. We now prove that the set $TS(f)$ is indeed a teaching set that uniquely specifies f within all of $\mathcal{D}_{\leq s}$.

We first observe that any term compatible with $TS(f)$ can only cover positive examples from one cluster of ϕ.

Lemma 6. *Let y be any positive example in $TS(f)$ and let T be any term that covers y and is compatible with $TS(f)$. Let C be the cluster of ϕ that covers y. Then if z is any positive example in $TS(f)$ that is not covered by C, T does not cover z.*

Proof. If C satisfies condition (i) then y must be a cogent corner point and Lemma 3 gives the desired conclusion. If C does not satisfy (i) but satisfies (ii), then the conclusion follows from Lemma 5. $\square$

The next two lemmas show that any set of terms that covers the positive examples of a given cluster must precisely compute the entire cluster and only the cluster of the original function:

Lemma 7. *Let C be any case (i) cluster of DNF-size k. Let P_C be the intersection of the positive examples in $TS(f)$ with C. Let $\widehat{T}_1, \ldots, \widehat{T}_j$ be any set of $j \leq k$ terms such that the DNF $\widehat{T}_1 \vee \cdots \vee \widehat{T}_j$ both: (a) is compatible with $TS(f)$, and (b) covers every point in P_C. Then it must be the case that $j = k$ and $\widehat{T}_1 \vee \cdots \vee \widehat{T}_j$ exactly computes C (in fact each term $\widehat{T}_i$ is equivalent to T_i up to reordering).*

Proof. By Lemma 3, a term $\widehat{T}$ that covers a cogent antipodal point from term T_i cannot cover any of the other $2k - 2$ cogent antipodal points from other terms, and thus we must have $j = k$ since fewer than k terms cannot cover all of P_C. Moreover, any term $\widehat{T}_i$ must cover a pair of antipodal points corresponding to a single term (which wlog we call T_i). For each antipodal pair corresponding to a term T_i, the covering term $\widehat{T}_i$ must be of size at least $|T_i|$, and since they are cogent antipodal points, the covering term cannot be any longer than $|T_i|$, so in fact we have that $\widehat{T}_i$ and T_i are identical. This proves the lemma. □

Lemma 8. *Let C be any case (ii) cluster. Let P_C be the intersection of the positive examples in $TS(f)$ with C. Let $\widehat{T}_1, \ldots, \widehat{T}_j$ be any set of $j \leq 2$ terms such that the DNF $\widehat{T}_1 \vee \cdots \vee \widehat{T}_j$ both: (a) is compatible with $TS(f)$, and (b) covers every point in P_C. Then it must be the case that $j = 2$ and $\widehat{T}_1 \vee \widehat{T}_2$ exactly computes C.*

Proof. The fact that $BTS(C)$ is a teaching set (for the exactly-2-term DNF corresponding to C, relative to $\mathcal{D}_{\leq 2}$) implies the desired result, since no single term or 2-term DNF not equivalent to C can be consistent with $BTS(C)$, and any DNF $\widehat{T}_1 \vee \cdots \vee \widehat{T}_j$ as specified in the lemma must be consistent with $BTS(C)$. □

The pieces are in place for us to prove our theorem:

Theorem 7. *For any $f \in \mathcal{S}$, the set $TS(f)$ uniquely specifies f within $\mathcal{D}_{\leq s}$.*

Proof. By Lemma 6, positive points from each cluster can only be covered by terms that do not include any positive points from other clusters. By Lemmas 7 and 8, for each cluster C, the minimum number of terms required to cover all positive points in the cluster (and still be compatible with $TS(f)$) is precisely the DNF-size of C. Since f is an exactly-s-term DNF, Lemma 4 implies that using more than DNF-size(C) many terms to cover all the positive points in any cluster C will "short-change" some other cluster and cause some positive point to be uncovered. Thus any at-most-s-term DNF ϕ that is consistent with $TS(f)$ must have the property that for each cluster C, at most DNF-size(C) of its terms cover the points in P_C; so by Lemmas 7 and 8, these terms exactly compute C, and thus ϕ must exactly compute f. □

4.3 Average-case teaching dimension of DNFs

In this section we will show that all but at most a $\frac{O(s)}{2^n}$ fraction of functions in $\mathcal{D}_{\leq s}$ are in fact in $\mathcal{S}$. We do this by showing that at least a $1 - \frac{O(s)}{2^n}$ fraction of functions in $\mathcal{D}_{\leq s}$ are in the easy-to-teach set $\mathcal{S}$, i.e. they belong to $\mathcal{D}_s$ and are such that each cluster satisfies either condition (i) or (ii) from Section 4.2. Since we have shown that each $f \in \mathcal{S}$ can be uniquely specified within $\mathcal{D}_{\leq s}$ using $O(ns)$ examples, this will easily yield that the average teaching number over all of $\mathcal{D}_{\leq s}$ is $O(ns)$.

First we show that most functions in $\mathcal{D}_{\leq s}$ are in fact in $\mathcal{D}_s$. We can bound $|\mathcal{D}_i|$ using the same approach as we did for monotone DNFs.

Lemma 9. *For $i < (9/7)^{n/3}$, we have $\frac{1}{2} \cdot \frac{3^{ni}}{i!} \leq |\mathcal{D}_i| \leq \frac{3^{ni}}{i!}$.*

Proof. As in Lemma 2, the upper bound is easy; we may bound the number of functions in $\mathcal{D}_i$ by the number of ways to choose i terms from the set of all 3^n possible terms over variables $x_1, \dots, x_n$. This is $\binom{3^n}{i} \leq \frac{3^{ni}}{i!}$.

For the lower bound, we first note that a DNF formula consisting of i terms that are all pairwise far from each other cannot be logically equivalent to any other DNF over a different set of i terms. We will show that at least half of all 3^{ni} possible sequences of i terms have the property that all i terms in the sequence are pairwise far from each other; this gives the lower bound (since each such set of i terms can be ordered in $i!$ different ways).

So consider a uniform random draw of i terms $T_1, \dots, T_i$ from the set of all 3^n possible terms. The probability that T_1 and T_2 are close is the probability that they have no strong differences plus the probability that they have exactly one strong difference. This is $(7/9)^n + n(7/9)^{n-1}(2/9) < (n+1)(7/9)^n$. By a union bound over all pairs of terms, the probability that any pair of terms is close at most $\binom{i}{2}(n+1)(7/9)^n$ which is less than $1/2$ for $i < (9/7)^{n/3}$. □

As in Section 3, as a corollary we have that $\frac{|\mathcal{D}_s|}{|\mathcal{D}_{\leq s-1}|} \geq \frac{3^n}{4s}$ for $s \leq (9/7)^{n/3}$.

We now bound the number of DNFs in $\mathcal{D}_s$ that are not in $\mathcal{S}$. To do this, we consider choosing s terms at random with replacement from all 3^n terms:

Lemma 10. *Fix any $s \leq (9/8)^{n/25}$. Let $f = T_1, \dots, T_s$ be a sequence of exactly s terms selected by independently choosing each T_i uniformly from the set of all 3^n possible terms. Let $A(T_i)$ denote the event that term T_i in f has no cogent antipodal pairs, and $B(T_i)$ denote the event that there is more than one other term close to T_i in f. Then $\Pr[\exists T_i \in f : A(T_i) \& B(T_i)] \leq \frac{O(s)}{2^n}$, where the probability is taken over the choice of f.*

Using Lemma 10 we can bound the number of functions $f \in \mathcal{D}_s$ that are not in $\mathcal{S}$. If $f \in \mathcal{D}_s \setminus \mathcal{S}$, then f must have a DNF formula representation $\phi = T_1 \vee \cdots \vee T_s$ in which some term T_i:

1. has no cogent antipodal pairs, and
2. has at least two other terms T_j, T_k that are close to it.

(If there were no such term, then for any representation $\phi = T_1 \vee \cdots \vee T_s$ for the function f, every T_i is contained in either a cluster of DNF-size 1 or 2, or a cluster of DNF-size k with a pair of good antipodal points around it. But then ϕ would be in $\mathcal{S}$.)

We will call such a syntactic DNF formula "bad." Lemma 10 tells us that the number of bad syntactic formulas is at most $\frac{3^{ns}O(s)}{2^n}$, since there are 3^{ns} syntactic formulas. Notice that any bad formula ϕ must have s distinct terms (since the function it computes belongs to $\mathcal{D}_s$), and since these terms can be ordered in $s!$ different ways, there are at least $s!$ bad formulas that compute the same function as ϕ. Consequently the number of bad functions in $\mathcal{D}_s$, $|\mathcal{D}_s \setminus \mathcal{S}|$, is at most $\frac{O(s)}{2^n}\frac{3^{ns}}{s!}$. By Lemma 9, $|\mathcal{D}_s|$ is at at least $\frac{3^{ns}}{2s!}$. This gives the following:

Corollary 2. $\frac{|\mathcal{D}_s \setminus \mathcal{S}|}{|\mathcal{D}_s|} \leq \frac{O(s)}{2^n}$.

We now proceed to prove Lemma 10.

Proof. The bulk of the argument is in showing that $\Pr[A(T_1) \ \& \ B(T_1)]$ is at most $O(1) \cdot 2^{-n}$; once this is shown a union bound gives the final result.

We condition on the outcome of T_1. Using the fact that each variable occurs independently in T_1 (either positive or negated) with probability 2/3, a Chernoff bound gives that $\Pr[|T_1| < .08n] \leq 2^{-n}$, so we have that

$$\Pr[A(T_1) \ \& \ B(T_1)] \leq 2^{-n} + \sum_{\mathcal{T}:|\mathcal{T}| \geq .08n} \Pr[A(T_1) \ \& \ B(T_1) \mid (T_1 = \mathcal{T})] \cdot \Pr[T_1 = \mathcal{T}].$$

Next we show that $\Pr[A(T_1) \ \& \ B(T_1) \mid (T_1 = \mathcal{T})] \leq O(1) \cdot 2^{-n}$ for every $\mathcal{T}$ satisfying $|\mathcal{T}| \geq .08n$; this implies an $O(1) \cdot 2^{-n}$ bound on $\Pr[A(T_1) \ \& \ B(T_1)]$. To do this we consider a third event which we denote by $C(T_1)$; this is the event that T_1 is close to at most 25 of the terms $T_2, \ldots, T_s$. Clearly we have that

$$\begin{aligned} \Pr[A(T_1) \ \& \ B(T_1) \mid (T_1 = \mathcal{T})] &= \Pr[A(T_1) \ \& \ B(T_1) \ \& \ \neg C(T_1) \mid (T_1 = \mathcal{T})] \\ &\quad + \Pr[A(T_1) \ \& \ B(T_1) \ \& \ C(T_1) \mid (T_1 = \mathcal{T})] \quad (1) \end{aligned}$$

and we proceed by bounding each of the terms in (1).

The first term is at most $\Pr[\neg C(T_1) \mid (T_1 = \mathcal{T})]$. Fix any $\alpha \in [.08, 1]$ and any term $\mathcal{T}$ of length αn, and fix $T_1 = \mathcal{T}$. Then the probability (over a random draw of T_2 as in the statement of the lemma) that T_2 is close to T_1 is the probability that T_1 and T_2 have one strong difference plus the probability that T_1 and T_2 have no strong difference, which is exactly $\alpha n \frac{1}{3}\left(\frac{2}{3}\right)^{\alpha n - 1} + \left(\frac{2}{3}\right)^{\alpha n} \leq 2\alpha n \left(\frac{2}{3}\right)^{\alpha n}$. Using the independence of the terms $T_2, \ldots, T_s$ and a union bound, it follows that the probability that there exists any set of K terms in f which are all close to T_1 is at most $\binom{s}{K}(2\alpha n)^K \left(\frac{2}{3}\right)^{K\alpha n}$. It is not hard to verify that for any $1 \leq s \leq (9/8)^{n/25}$, any $K \geq 26$, and any $\alpha \in [.08, 1]$, this quantity is asymptotically less than 2^{-n}.

It remains to bound the second term of (1) by $O(1) \cdot 2^{-n}$. We do this using the following observation:

Proposition 1. *Let $f = T_1, \ldots, T_s$ be any sequence of s terms. If T_1 has no cogent antipodal pairs with respect to f and is close to at most K of the terms $T_2, \ldots, T_s$, then there must be some term among $T_2, \ldots, T_s$ that is close to T_1 and contains at most $k = \lceil \log K \rceil + 1$ variables not already in T_1.*

Proof. We show that if every term in f close to T_1 contains more than k variables, there must remain some cogent antipodal pair for T_1. Let r be the number of variables in T_1 and let $\ell = n - r$. For any $z \in \{0,1\}^\ell$ let $Q_{T_1}(z)$ denote the set of points in $\{0,1\}^n$ consisting of the antipodal pair induced by z on T_1 (these two points each satisfy T_1) and the $2r$ neighbors of these points that do not satisfy T_1. Thus $Q_{T_1}(z) = Q_{T_1}(\overline{z})$, and there are $2^{\ell-1}$ distinct $Q_{T_1}(z)$, each representing a possible cogent antipodal pair.

Consider a term T_i that is close to T_1, and partition its satisfying assignments according to the 2^ℓ assignments on the ℓ variables not contained in T_1. Since T_i will only eliminate the cogent antipodal pair represented by the neighborhood $Q_{T_1}(z)$ if it covers some point in $Q_{T_1}(z)$, T_i can only eliminate as many cogent antipodal pairs as it has partitions. But if T_i contains more than k of the ℓ variables not already in T_1, then there are fewer than $2^{\ell-k}$ different ways to set the ℓ bits outside of T_1 to construct a satisfying assignment for T_i, and T_i has fewer than $2^{\ell-k}$ different partitions. Since by assumption there are at most $K \leq 2^{k-1}$ terms close to T_1, there are fewer than $2^{k-1} \cdot 2^{\ell-k} = 2^{\ell-1}$ different $Q_T(z)$ eliminated, and T must have a cogent antipodal pair left. □

By Proposition 1, we know that if $A(T_1)$ occurs (T_1 has no cogent antipodal pairs) and $C(T_1)$ occurs (T_1 is close to no more than $K = 25$ other terms), then there must be some term close to T_1 that has at most $k = 6$ variables not in T_1. Thus we have that $\Pr[A(T_1)\ \&\ B(T_1)\ \&\ C(T_1) \mid (T_1 = \mathcal{T})]$ is at most the probability there exist two terms close to T_1, one of which contains at most $k = 6$ variables not in T_1. We saw earlier that the probability that a randomly chosen term is close to T_1 is at most $2\alpha n(2/3)^{\alpha n}$. However, the probability that a randomly chosen term is close to T_1 *and* contains at most 6 variables not in T_1 is much lower (because almost all of the $(1-\alpha)n$ variables not in T_1 are constrained to be absent from the term); more precisely this probability is at most $2\alpha n \binom{(1-\alpha)n}{6} \left(\frac{2}{3}\right)^{\alpha n} \left(\frac{1}{3}\right)^{(1-\alpha)n-6}$. A union bound over all possible pairs of terms gives us that the second term of (1) is at most $2\alpha n \binom{s}{2} \binom{(1-\alpha)n}{6} 3^6 \left(\frac{2}{3}\right)^{2\alpha n} \left(\frac{1}{3}\right)^{(1-\alpha)n}$. It is straightforward to check that this is at most $O(1) \cdot 2^{-n}$ for all $1 \leq s \leq (9/8)^{n/25}$ and all $\alpha \in [0,1]$.

Thus, we have bounded $\Pr[A(T_1)\ \&\ B(T_1)]$ by $O(1) \cdot 2^{-n}$. A union bound over the s terms gives that $\Pr[\exists T_i \in f : A(T_i)\ \&\ B(T_i)]$ is at most $O(s)2^{-n}$, and the lemma is proved. □

Theorem 8. *Let $s \leq (9/8)^{n/25}$. The average teaching dimension of $\mathcal{D}_{\leq s}$, the class of DNFs over n variables with at most s terms, is $O(ns)$.*

Proof. Theorem 7 gives us that the teaching number of any concept in $\mathcal{S} \subset \mathcal{D}_s$ is $O(ns)$. By Lemma 9, we have that $|\mathcal{D}_{\leq s-1}| \leq \frac{4s}{3^n} |\mathcal{D}_s|$. This leaves us with $\mathcal{D}_s \setminus \mathcal{S}$, whose size we bounded by $\frac{O(s)}{2^n}|\mathcal{D}_s|$ in Corollary 2. Combining these bounds, we are ready to bound the average teaching number of $|\mathcal{D}_{\leq s}|$. Since we can teach any bad concept with at most 2^n examples, the average teaching number is at most

$$\frac{O(ns)|\mathcal{S}| + 2^n(|\mathcal{D}_{\leq s-1}| + |\mathcal{D}_s \setminus \mathcal{S}|)}{|\mathcal{D}_s| + |\mathcal{D}_{\leq s-1}|} \leq \frac{O(ns)|\mathcal{D}_s| + 2^n(\frac{4s}{3^n}|\mathcal{D}_s| + \frac{O(s)}{2^n}|\mathcal{D}_s|) +}{|\mathcal{D}_s| + |\mathcal{D}_{\leq s-1}|}$$
$$\leq O(ns) + (2/3)^n \cdot 4s + O(s) = O(ns)$$

and the theorem is proved. □

As in Corollary 1, we have $2^n \leq \text{poly}(s)$ if $s > (9/8)^{n/25}$, and thus the worst-case teaching number 2^n is actually $\text{poly}(n, s)$ for such large s. This gives the following corollary:

Corollary 3. *Let s be any value $1 \leq s \leq 2^n$. The class $\mathcal{D}_{\leq s}$ of at-most-s-term DNF has average teaching number poly(n, s).*

5 Teaching Dimension of k-juntas

A Boolean function f over n variables depends on its i-th variable if there are two inputs $x, x' \in \{0,1\}^n$ that differ only in the i-th coordinate and that have $f(x) \neq f(x')$. A *k-junta* is a Boolean function which depends on at most k of its n input variables. The class of k-juntas (or equivalently NC^0_k functions) is well studied in computational learning theory, see *e.g.,* [4, 11, 1]. We write $\mathcal{J}_k$ to denote the class of Boolean functions $f \colon \{0,1\}^n \to \{0,1\}$ that depend on exactly k variables, and we write $\mathcal{J}_{\leq k}$ to denote the class $\mathcal{J}_{\leq k} = \cup_{k' \leq k} \mathcal{J}_{k'}$ of Boolean functions over $\{0,1\}^n$ that depend on at most k variables, *i.e.,* $\mathcal{J}_{\leq k}$ is the class of all k-juntas.

We analyze the worst-case and average-case teaching dimensions of the class of k-juntas, and show that while the worst-case teaching dimension has a logarithmic dependence on n, the average-case dimension has no dependence on n. Thus k-juntas are another natural concept class where there is a substantial asymptotic difference between the worst-case and average teaching dimensions.

Worst-Case teaching dimension of k-juntas. We recall the following:

Definition 1. *Let $k \leq n$. A set $S \subseteq \{0,1\}^n$ is said to be an (n,k)-*universal set *if for any $1 \leq i_1 < i_2 \ldots < i_k \leq n$, it holds that $\forall y \in \{0,1\}^k, \exists x \in S$ satisfying $(x_{i_1}, \ldots, x_{i_k}) = (y_1, \ldots, y_k)$*

Nearly matching upper and lower bounds are known for the size of (n,k)-universal sets:

Theorem 9 ([14]). *Let $k \leq n$. Any (n,k)-universal set has size $\Omega(2^k \log n)$, and there exists an (n,k)-universal set of size $O(k2^k \log n)$.*

This straightforwardly yields:

Theorem 10. *The teaching dimension of $\mathcal{J}_{\le k}$ is at least $\Omega(2^k \log n)$ and at most $O(k2^k \log n)$.*

Proof. For the lower bound, we show that any teaching set for the identically-0 concept $c \equiv \mathbf{0}$ (which is a k-junta for any $k \ge 0$) must be an (n,k)-universal set. Suppose $S \subseteq \{0,1\}^n$ is not an (n,k)-universal set, *i.e.,* there is some $i_1 < \cdots < i_k$ and some $y \in \{0,1\}^k$ such that for every $x \in S$ we have $(x_{i_1}, \ldots, x_{i_k}) \ne (y_1 \ldots y_k)$. Then the k-junta defined as

$$c'(x) = \begin{cases} 1 & \text{if } (x_{i_1} \ldots x_{i_k}) = (y_1 \ldots y_k) \\ 0 & \text{otherwise} \end{cases}$$

labels S the same way as c.

Now we prove the upper bound. Let c be any k-junta that has $R = \{i_1, \ldots, i_r\}$ as its set of relevant variables (so $r \le k$). We describe a teaching set for c. For each relevant variable $i_j \in R$, there is a pair of examples $x, x' \in \{0,1\}^n$ that disagree only in their i_j-th bit and have $c(x) \ne c(x')$. Let the set S consists of these $2r$ examples together with an (n,k)-universal set; we will argue that S is a teaching set for c and thus prove the theorem.

Suppose that c' is some k-junta that is consistent with S. Clearly c' must depend on every variable in R or else it would label one of the first $2r$ examples differently from c. We claim that c' cannot depend on any additional variables. Suppose to the contrary that c' depends on exactly q additional variables $j_1, \ldots, j_q$. Given $a \in \{0,1\}^r$ and $b \in \{0,1\}^q$, let $V(a,b) = \{x \in \{0,1\}^n : (x_{i_1} \ldots x_{i_r}) = a$ and $(x_{j_1} \ldots x_{j_q}) = b\}$. Since c' depends on $j_1, \ldots, j_q$, there must be some $a \in \{0,1\}^r$ and $b \ne b' \in \{0,1\}^q$ such that all the examples in $V(a,b)$ take one value under c' while all the examples in $V(a,b')$ take the other value under c'. Furthermore, since S is an (n,k)-universal set and $|a|+|b| \le k$, S must contain some example x^1 from $V(a,b)$ and some example x^2 from $V(a,b')$. But c only depends on variables $i_1, \ldots, i_r$, so c assigns x^1 and x^2 the same label while c' does not. Thus c and c' must have the exact same set of $r \le k$ relevant variables. Since they agree on an (n,k)-universal set, they agree for every setting of those r variables, and thus they agree on all of $\{0,1\}^n$. □

Average-case teaching dimension of k-juntas. The idea is similar to the case of monotone DNF: we show that k-juntas with exactly k relevant variables can be taught with 2^k examples (independent of n), and then use the fact that the overwhelming majority of k-juntas have exactly k relevant variables.

Lemma 11. *Let c be any concept in $\mathcal{J}_k$. Then the teaching dimension of c with respect to $\mathcal{J}_{\le k}$ is at most 2^k.*

Proof. Given any k-junta c with exactly k relevant variables, let S be the set of 2^k examples in which all irrelevant variables are always set to 0 and the relevant variables range over all 2^k possible settings. It is straightforward to see that S is a teaching set for c. □

We now claim that $\frac{1}{2}\binom{n}{k}2^{2^k} \leq |\mathcal{J}_{\leq k}| \leq \binom{n}{k}2^{2^k}$. The upper bound is clear since any k-junta can be specified by presenting k variables ($\binom{n}{k}$ possibilities) and a Boolean function on those k variables (2^{2^k} possibilities). The lower bound (which is very crude but sufficient for our purposes) follows from the easily verified fact that at least half of all 2^{2^k} functions on $\{0,1\}^k$ in fact depend on all k variables. It is easy to see from these bounds that $|\mathcal{J}_k|$ strictly increases with k for all k, and thus we have $|\mathcal{J}_{\leq k-1}| \leq (k-1)|\mathcal{J}_{k-1}| \leq (k-1)\binom{n}{k-1}2^{2^{k-1}}$.

By Lemma 11 we can specify any function in $\mathcal{J}_k$ with at most 2^k examples, and by Theorem 9 we can specify any of the other functions in $\mathcal{J}_{\leq k}$ (*i.e.,* any function in $\mathcal{J}_{\leq k-1}$) with at most $O(k2^k \log n)$ many examples. It follows that the average teaching dimension of $\mathcal{J}_{\leq k}$ is at most

$$\frac{2^k|\mathcal{J}_k| + O(k2^k \log n) \cdot |\mathcal{J}_{\leq k-1}|}{|\mathcal{J}_k| + |\mathcal{J}_{\leq k-1}|} \leq 2^k + \frac{O(k2^k \log n) \cdot (k-1)\binom{n}{k-1}2^{2^{k-1}}}{\frac{1}{2}\binom{n}{k}2^{2^k}}.$$

The second term on the right simplifies to

$$\frac{O(k2^k \log n) \cdot k(k-1)}{2^{2^{k-1}}(n-k+1)}$$

which is easily seen to be $o(1)$ for any k. We have thus proved:

Theorem 11. *The average teaching dimension of the class $\mathcal{J}_{\leq k}$ of k-juntas is at most $2^k + o(1)$.*

6 Sparse GF_2 Polynomials

A GF_2 *polynomial* is a multilinear polynomial with 0/1 coefficients that maps $\{0,1\}^n$ to $\{0,1\}$ where all arithmetic is done modulo 2. Since addition mod 2 corresponds to parity and multiplication corresponds to AND, a GF_2 polynomial can be viewed as a parity of monotone conjunctions. It is well known, and not hard to show, that every Boolean function $f\colon \{0,1\}^n \to \{0,1\}$ has a unique GF_2 polynomial representation. (For example, the parity function has $x_1 \oplus \cdots \oplus x_n$ as its GF_2 polynomial, and $x_1 \vee x_2$ has $x_1 \oplus x_2 \oplus x_1x_2$.)

A natural measure of the size of a GF_2 polynomial is the number of monomials that it contains. In keeping with our usual notation, let $\mathcal{G}_s$ denote the class of all Boolean functions $f\colon \{0,1\}^n \to \{0,1\}$ that have GF_2 polynomial representations with exactly s monomials and let $\mathcal{G}_{\leq s}$ denote $\cup_{s' \leq s}\mathcal{G}_{s'}$. We sometimes refer to functions in $\mathcal{G}_{\leq s}$ as being *s-sparse GF_2* polynomials. The class of s-sparse GF_2 polynomials has been studied by several researchers in learning theory and complexity theory, see *e.g.,* [12, 5, 13].

Roth and Benedek [12] showed that any $f \in \mathcal{G}_{\leq s}$ is uniquely determined by the values it assumes on those $x \in \{0,1\}^n$ that contain at least $n-(1+\lfloor \log_2 s \rfloor)$ many 1s. They also showed that it is in fact necessary to specify the value of f on every such point even in order to uniquely determine the parity (even or odd) of $|f^{-1}(1)|$ where f ranges over all of $\mathcal{G}_{\leq s}$. We thus have:

Theorem 12 ([12]). *Fix any $1 \leq s \leq 2^n$. The (worst-case) teaching dimension of $\mathcal{G}_{\leq s}$ is $\sum_{i=0}^{1+\lfloor \log_2 s \rfloor} \binom{n}{i}$ (which is $n^{\Theta(\log s)}$ for s subexponential in n).*

In contrast, we show in the next subsection that if s is sufficiently small, the average-case teaching dimension of $\mathcal{G}_{\leq s}$ is $O(ns)$:

Theorem 13. *Fix $1 \leq s \leq (1-\epsilon)\log_2 n$, where $\epsilon > 0$ is any constant. Then the average-case teaching dimension of $\mathcal{G}_{\leq s}$ is at most $ns + 2s$.*

For $s = \omega(1)$, $s < (1-\epsilon)\log_2 n$, this gives a superpolynomial separation between worst-case and average-case teaching dimension of s-sparse GF_2 polynomials.

Proof of Theorem 13. We now define the "nice" (easy-to-teach) subset of $\mathcal{G}_{\leq s}$, in analogy with $\mathcal{S}$ in Section 4. We say that a function $f = M_1 \oplus \cdots \oplus M_s \in \mathcal{G}_s$ is *individuated* if for each $i = 1, \ldots, s$ there is some $j \in \{1, \ldots, n\}$ such that the variable x_j occurs in monomial M_i and does not occur in any of the other $s-1$ monomials. Let $\mathcal{I} \subseteq \mathcal{G}_s$ denote the set of all functions in $\mathcal{G}_s$ that are individuated.

We first show that any function in $\mathcal{I}$ can be specified using few examples:

Lemma 12. *For any $f \in \mathcal{I}$, the teaching dimension of f with respect to $\mathcal{G}_{\leq s}$ is at most $ns + 2s - 1$.*

Proof. We introduce some useful terminology. Given $x^1, \ldots, x^r \in \{0,1\}^n$, we write $\text{join}(x^1, \ldots, x^r)$ to denote the string $z \in \{0,1\}^n$ that has for all $i = 1, \ldots, n$, $z_i = \max\{x_i^1, \ldots, x_i^r\}$.

Let $f = M_1 \oplus \cdots \oplus M_s \in \mathcal{I}$ be any individuated GF_2 polynomial. For $i = 1, \ldots, s$ let y^i denote the minimal (with respect to bitwise $\leq$ ordering described above) assignment that satisfies M_i, *i.e.,* y^i has 1s in precisely the variables contained in M_i. Note that since f is individuated the points $y^1, \ldots, y^s$ are all pairwise incomparable w.r.t. the bitwise partial ordering. Thus we have $f(y^i) = 1$ but $f(x) = 0$ for any x such that $x < y^i$ for some i. We sometimes say that y is above x if $x \leq y$.

Let $S \subset \{0,1\}^n$ be the set which contains: (a) each y^i (which is a positive example) and all of its neighbors that can be obtained by flipping a single 1 to 0 (all of these are negative examples); and (b) the $(s-1)$ additional points $z^2 = \text{join}(y^1, y^2)$, $z^3 = \text{join}(y^1, y^2, y^3)$, $\ldots$, $z^s = \text{join}(y^1, y^2, \ldots, y^s)$ (it is not hard to see that z^i is a positive example for i odd and a negative example for i even, since z^i satisfies precisely the monomials $M_1, \ldots, M_i$ and $M_1 \oplus \cdots \oplus M_s$ is individuated). There are at most $(n+1)s$ points from (a) and $s-1$ points from (b) so we have $|S| \leq ns + 2s - 1$.

We will show that S is a teaching set for f and thus prove the lemma. So suppose that $\hat{f} = \widehat{M}_1 \oplus \cdots \oplus \widehat{M}_r$ is some GF_2 polynomial that is consistent with S where $r \leq s$. Let us write $\hat{y}^j$ for the minimal assignment that satisfies $\widehat{M}_j$.

We first observe that since y^1 is a positive example, there must be at least one $\hat{y}^j$ such that $\hat{y}^j \leq y^1$. Since $y^1 < z^2$ and z^2 is a negative example, there must be at least two $\hat{y}^j$ such that $\hat{y}^j \leq z^2$. Since the labels of $z^2, z^3, \ldots$ always alternate, proceeding in this fashion there must be at least s many $\hat{y}^j$ such that $\hat{y}^j \leq z^s$. It follows that $r = s$, that y^1 is above precisely one $\hat{y}^1$, and that in fact each z^i is above precisely i of the $\hat{y}^j$'s (call them $\hat{y}^1, \ldots, \hat{y}^i$).

Now the negative examples below y^1 show that in fact we must have $\hat{y}^1 = y^1$. Since z^2 is above exactly one other $\hat{y}^j$ besides $\hat{y}^1$ (namely $\hat{y}^2$), and it is above y^2 which is labeled positive, we must have $\hat{y}^2 \leq y^2$; but since all of y^2's downward neighbors are labeled negative, it must be the case that $\hat{y}^2 = y^2$. Similar logic applied successively to $z^3, \ldots, z^s$ shows that each of $\hat{y}^3, \ldots, \hat{y}^s$ must equal the corresponding $y^3, \ldots, y^s$. Thus we have $\widehat{M_i} = M_i$ for $i = 1, \ldots, s$, so $\hat{f} = f$ and the lemma is proved. □

Now observe that $|\mathcal{G}_s| = \binom{2^n}{s} < \frac{2^{ns}}{s!}$, and thus $(\frac{2^n}{s})^s \leq |\mathcal{G}_{\leq s}| = |\mathcal{G}_s| + |\mathcal{G}_{\leq s-1}| < \frac{2^{ns}}{s!} + (s-1)\frac{2^{ns-n}}{(s-1)!} = \frac{2^{ns}}{s!} + \frac{2^{ns-n}}{(s-2)!}$. Our next lemma shows that almost every function in $\mathcal{G}_s$ (and thus almost every function in $\mathcal{G}_{\leq s}$) is in fact individuated:

Lemma 13. *We have* $|\mathcal{I}| \geq \frac{2^{ns}}{s!}(1 - s \cdot e^{-n^\epsilon})$*, and thus there are at most* $s \cdot e^{-n^\epsilon} \cdot \frac{2^{ns}}{s!} + \frac{2^{ns-n}}{(s-2)!}$ *many functions in* $\mathcal{G}_{\leq s} \setminus \mathcal{I}$.

Proof. Let $(M_1, \ldots, M_s)$ be a sequence of s monomials obtained by drawing each one uniformly from all 2^n possible monomials. We will show that of the 2^{ns} possible outcomes for $(M_1, \ldots, M_s)$, at most an $s \cdot e^{-n^\epsilon}$ fraction have the property that the corresponding GF_2 polynomial $M_1 \oplus \cdots \oplus M_s$ is not individuated, and consequently the number of sequences for which the corresponding GF_2 polynomial *is* individuated is at least $2^{ns}(1 - s \cdot e^{-n^\epsilon})$. Each such sequence clearly consists of s distinct monomials (since no sequence in which some monomial occurs more than once can be individuated), so accounting for the $s!$ different orderings of s distinct elements, we have that there are at least $2^{ns}(1 - s \cdot e^{-n^\epsilon})/s!$ many individuated GF_2 polynomials.

We say that a variable individuates a monomial M_i if it occurs in M_i but in no other M_j. For any fixed variable x_j, and fixed index $1 \leq i \leq s$, the probability (over the random choice of $(M_1, \ldots, M_s)$) that x_j individuates M_i is precisely $1/2^s$, since x_j must occur in M_i (probability $1/2$) and must be absent from each of the other $s-1$ terms (probability $1/2^{s-1}$). By independence, the probability that none of the n variables individuates M_i is $\left(1 - \frac{1}{2^s}\right)^n \leq e^{-n/2^s} \leq e^{-n^\epsilon}$, where we have used the fact that $s \leq (1-\epsilon)\log_2 n$. A union bound now gives that the probability that any of the s monomials $M_1, \ldots, M_s$ is not individuated by any variable is at most $s \cdot e^{-n^\epsilon}$. □

By Lemma 12 we can specify any function in $\mathcal{I}$ with at most N examples, and by Theorem 12 we can specify any of the other functions in $\mathcal{G}_{\leq s}$ with at most $n^{O(\log s)}$ many examples. It follows from Lemma 13 that the average teaching dimension of $\mathcal{G}_{\leq s}$ is at most

$$\frac{N|\mathcal{I}| + n^{O(\log s)} \cdot |\mathcal{G}_{\leq s} \setminus \mathcal{I}|}{|\mathcal{G}_{\leq s}|} \leq N + \frac{n^{O(\log s)} \cdot (s \cdot e^{-n^\epsilon} \cdot \frac{2^{ns}}{s!} + \frac{2^{ns-n}}{(s-2)!})}{(\frac{2^n}{s})^s}.$$

The second term on the right simplifies to $s^s \cdot n^{O(\log s)} \cdot (s \cdot e^{-n^\epsilon}/s! + 2^{-n}/(s-2)!)$, which is easily seen to be $o(1)$ since ϵ is a constant greater than 0 and $s \le (1-\epsilon)\log n$. This proves Theorem 13. □

While our proof technique does not extend to s that are larger than $\log n$, it is possible that different methods could establish a poly(n,s) upper bound on average teaching dimension for the class $\mathcal{G}_{\le s}$ of s-sparse GF_2 polynomials for a much larger range of values of s. This is an interesting goal for future work.

References

1. M. Alekhnovich, M. Braverman, V. Feldman, A. Klivans, and T. Pitassi. Learnability and automatizability. In *Proceedings of the 45th IEEE Symposium on Foundations of Computer Science*, pages 621–630, 2004.
2. M. Anthony, G. Brightwell, and J. Shawe-Taylor. On specifying Boolean functions by labelled examples. *Discrete Applied Math.*, 61(1):1–25, 1995.
3. F. Balbach. Teaching classes with high teaching dimension using few examples. In *Proc. 18th Annual COLT*, pages 637–651, 2005.
4. A. Blum. Learning a function of r relevant variables (open problem). In *Proc. 16th Annual COLT*, pages 731–733, 2003.
5. N. Bshouty and Y. Mansour. Simple Learning Algorithms for Decision Trees and Multivariate Polynomials. *SIAM J. Comput.*, 31(6):1909–1925, 2002.
6. J. Cherniavsky and R. Statman. Testing: An abstract approach. In *Proceedings of the 2nd Workshop on Software Testing*, 1998.
7. S. Goldman and M. Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50(1):20–31, February 1992.
8. S. Goldman, R. Rivest, and R. Schapire. Learning binary relations and total orders. *SIAM Journal on Computing*, 22(5):1006–1034, October 1993.
9. Christian Kuhlmann. On teaching and learning intersection-closed concept classes. In *Proc. 4th EUROCOLT*, pages 168–182, 1999.
10. E. Kushilevitz, N. Linial, Y. Rabinovich, and M. Saks. Witness sets for families of binary vectors. *J. Combinatorial Theory*, 73(2):376–380, 1996.
11. E. Mossel, R. O'Donnell, and R. Servedio. Learning functions of k relevant variables. *J. Comput. & Syst. Sci.*, 69(3):421–434, 2004.
12. R. Roth and G. Benedek. Interpolation and approximation of sparse multivariate polynomials over $GF(2)$. *SIAM J. Comput.*, 20(2):291–314, 1991.
13. R. Schapire and L. Sellie. Learning sparse multivariate polynomials over a field with queries and counterexamples. *J. Comput. & Syst. Sci.*, 52(2):201–213, 1996.
14. Gadiel Seroussi and Nader Bshouty. Vector sets for exhaustive testing of logic circuits. *IEEE Trans. on Information Theory*, 34(3):513–522, 1988.